Research on Examinations in India

Research on Examinations in India

A Edwin Harper, Jr and Vidya Sagar Misra
with the assistance of
Ramji Lal Gupta

MLSU - CENTRAL LIBRARY
81753CL

NATIONAL COUNCIL OF EDUCATIONAL RESEARCH
AND TRAINING

January 1976

Pausa 1897

P D 1T

© *National Council of Educational Research and Training, 1976*

This research project was
supported by a grant from
National Council of Educational
Research and Training
New Delhi

Rs 30 10 $ 9 00 £ 3 00

PUBLISHED AT THE PUBLICATION DEPARTMENT BY SMT J ANJANI DAYANAND, SECRETARY, NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING, SRI AUROBINDO MARG, NEW DELHI 110016 AND PRINTED AT THE STATESMAN PRESS, 4, CHOWRINGHEE SQUARE CALCUTTA

Foreword

PUBLIC EXAMINATIONS in India play a pivotal role in determining the functional content and method of education They carry a good deal of prestige and act as a passport for admission to higher courses as also for entering jobs The agencies conducting public examinations are quite conscious of their responsibility and in their own way take various steps to ensure that they are run as efficiently as possible Adequate precautions are taken in the appointment of paper-setters, in instructions to examiners, etc , in order to minimise the factor of unreliability

Nevertheless, studies conducted both in India and outside have brought out very clearly the limited value of such examinations Dr A E Harper's study is one such One major difficulty in conducting such studies in India has been the faith placed in external examinations and the element of secrecy attached to them Dr Harper has succeeded in overcoming this difficulty and his research is based on live data obtained in simulated conditions

He has studied a sample of answer scripts in four high school subjects by getting them examined by more than one examiner This massive study has brought out the numerous deficiencies inherent in external examinations His findings will come as an eye-opener both to those functioning at various levels in examining agencies and to policy-makers in the field of school education They will also be useful to decision makers and will bring home to them the fact that the conventional steps taken by the examining agencies are not adequate to overcome the evils of the system It is hoped that Dr Harper's findings will open up new vistas for reforms in the present system of external examinations.

RAIS AHMED
Director
National Council of Educational Research and Training

NEW DELHI
APRIL 1975

To The Reader

THIS REPORT is addressed to all who are interested in our present examination system, and in its reform

Some of you are highly trained experts, who can talk of sigmas, and standard errors of measurement, as easily as about the price of mangoes

Others of you are teachers, administrators, educators, students, parents You know nothing about statistics, and even the coefficient of correlation is an enigma to you It is to *you* that this chapter is especially addressed

You will find this report written on three levels This means that there is some repetition. But it also means that it will be readable and useful to a larger number than might otherwise have been possible

First, for the impatient, is a chapter of findings alone (Chapter II) This is just a series of statements, without presenting any evidence This is the chapter labelled "Briefly Summarized". You may also find this useful for quotation

Second, for those who are interested in the major findings, but not in the detailed procedures and statistical analyses, you will find the "Highlights" (Chapters III and IV) most useful These are non-technical and can be read, and understood, by any teacher, parent, or other interested person In addition to this, we suggest that you at least glance at the Tables and Figures in the rest of the book. These will give you a good overview of what we have found

Third, the rest of the book is for those who want all the details.

Part Two—Detailed Reports begins with Chapter V which reviews other research in this field, particularly that done in India. Because this has never been reviewed before, we believe the reader will be surprised to find how much has already been done. Chapter VI describes the experimental design, and Chapter VII which attempts to clarify terminology and concepts, will be of interest to those with a serious concern in this field. Detailed reports follow on history, Hindi, biology and mathematics examinations. Mathematics is placed last because it is popularly considered the most reliable of these four examinations. The reader who seriously expects this is in for a surprise. Finally, there are three chapters of recommendations, with an emphasis on the practical, "it can be done now" type.

Part Two is, of course, primarily for those with a knowledge of psychometric statistics. *But others will find it useful, also.* Try it, even if you are not an "expert". Every technical term has been explained briefly, when it is introduced. And there are, of course, many parts which require no "technical" knowledge whatsoever.

In Part Two, we are sure that you will want to read at least the section related to the subject you teach. Do not be worried if there are parts you do not understand—you will certainly find much of interest in the parts that you can understand quite clearly.

There is one more way to review this book, if you do not have time to read it all: look at the Index. Just read through the Index, find topics that interest you, and look them up. This way you can go through the book quite quickly, and yet not miss anything which is likely to be of importance to *you*. It would be best, however, if you read Chapters I and II first, before using the Index to skim the rest.

Finally, here are some of the questions to which you can seek answers in this book:

"How reliable are our traditional type of examinations?"

"What happens when 90 experienced examiners mark the same answer books?"

"If an examiner re-examines a batch of answer books, will he give them the same marks the second time?"

"How meaningful is a First Class?"

"Do experienced examiners agree on pass percentage for the same answer books?"

"What element does 'chance' play in passing, failing, and Divisions?"

"Will it help to have several examiners mark each answer book, and average the results ? Is there a cheaper way than this to make traditional-type examinations more reliable?"

"What improvements can be introduced without major changes in the present type and pattern of examinations?"

Responsibilities and Acknowledgements

The two research projects reported in this book were conducted under a grant from the National Council of Educational Research and Training, with the assistance of what was then the Central Evaluation Unit of the DEPSE, and the Teachers College, Columbia University Team However, the conclusions and opinions expressed are those of the authors who have signed this report, and are not to be construed as official statements of the sponsoring bodies

The senior author was Director of this research project He had working with him two Research Assistants One of these, Shri Vidya Sagar Misra, took major responsibilities at all stages of the research project, and has contributed far beyond the brief period during which he was on salary for the project Without his selfless service, his willingness to work far beyond normal hours (the computer was usually available after midnight), his intelligent and careful attention to detail, this research would have been quite impossible The second Research Assistant, Shri R L Gupta was a mathematician who contributed during the first five months of the research

The senior author is responsible for most of the writing of this report, although the junior author has read much of it and made many suggestions Chapter V, the Review of the Literature, however, is entirely the work of the junior author It was initially prepared as part of a Ph D thesis, and is included here to reach a wider audience and to provide relevant background for these studies

So much for responsibilities Acknowledgements are given with the

understanding that those acknowledged have contributed much that is worthwhile in this research, but are in no way responsible for any of its defects.

The planning for these research projects was carried out with extensive help from Dr. Walker Hill, Dr. R. H. Dave, and Dr. Harold Webster. Their continuing advice and guidance at all stages is gratefully acknowledged.

In the background, encouraging, facilitating, making official contacts, assisting in many subtle and essential ways were Dr. Paul Leonard, Shri L. S. Chandrakant, and Shri Raja Roy Singh. At crucial points, their support was invaluable.

Very grateful thanks are due to the Board of Secondary Education which provided the basic data, but which has asked to remain anonymous. (*Those who may already know the identity of this Board are asked to respect this request.*) As the Secretary pointed out, the results of this study are probably equally applicable to *any* Board—and it would be most unfortunate if partisan persons should try to use them to condemn a particular Board rather than the examination system itself. Nevertheless, let this stand as our acknowledgement for the very generous and wholehearted cooperation received from the Secretary of the Board of Secondary Education which furnished the answer books and names and addresses of experienced examiners.

The same reason (anonymity of the Board) prevents us from acknowledging by name the 134 experienced examiners who gave of their time and their skill to this research. For their interest in the project, their many illuminating comments, and their care in their work we are very grateful indeed. So, also, for the 4,010 anonymous students whose answer books we used.

Shri J. K Sen Gupta, Marketing Manager, Gaumount-Kalee Limited, Bombay supervised the reproduction of the answer books for the "ninety marking ten" project. In so doing, he pioneered new processes, never before available in India. Shri K. L. Gupta, Zonal Manager, L I C, Kanpur allowed his staff to do the card punching for us; Shri Raghawan, Section Officer assisted in obtaining IBM cards, and in other ways; and Shri R. Ramaswamy, Zonal Actuary, supervised the work.

Dr. H. K. Kesavan, Head, Computer Centre, Indian Institute of Technology, Kanpur, admitted the junior author to his computer programming course without charge, and was most generous in providing the computer time necessary for carrying through the "four thousand re-examined" project.

My personal thanks are due to Geonautics, Inc., of Washington, D C., for allowing me to use their LOCI-1 (a small desk-top computer)

in checking various computations My brother, Mr H MacDonald Harper, of Geonautics, programmed the computer and did some of the checking, and I am very grateful for his brotherly assistance

Statistical advice is essential in research of this sort Though they have also been acknowledged elsewhere, we repeat here the names of Doctors Harold Webster C R Rao, A B L Srivastava, and Helen Walker

Advice and critical comments on various drafts are invaluable in clarifying both issues and writing Dr Warren Findley, who pointed out what is good as well as what is weak or wrong, Dr Ram Gopal Misra, who raised many helpful questions my esteemed colleague Dr Sushil K V Liddle who advised on many aspects of these studies

My long suffering wife, Erika S Harper has contributed in more ways than I can possibly list She has provided work space kept me free from distractions, and served as a walking dictionary ("I didn t get my Ph D in spelling ', I tell my friends) Most important, she has read and commented on every word of the manuscript, and helped me to clarify and re-write many otherwise obscure points Her editorial help has done much to reduce the ivory tower technicalities to the reading level of the common man

Shri Jamuna Prasad Jaiswar has typed not only most of the drafts, but also the final manuscript, produced under deadline pressure with many hours of overtime work

A Edwin Harper, Jr

Allahabad
June 1970

Contents

FOREWORD v

TO THE READER vii

RESPONSIBILITIES AND ACKNOWLEDGEMENTS ix

PART ONE—HIGHLIGHTS

1 *Introduction* 1

2 *Briefly Summarized* 4

3 *Highlights of "Ninety Marking Ten'* 8

What happens when ninety experienced examiners mark the same ten history answer books ?—8
Are these differences just random errors ?—10
Can we trust Class or Division ?—11
Can we trust 'Pass Percentage' ?—12
Can examiners agree on who is best and who is worst ?—13
What element does 'chance' play in passing and failing ?—15
How reliable is this examination ?—16
Will it help to have several examiners mark each answer book, and average the results ?—17

4 *Highlights of "Four Thousand Re-Examined'* 19

Differences between examiners?—19
How large a mistake is likely to be made for any single candidate ?—20

Can we trust "Pass Percentage" ?—22
Can we trust "Class or Division" ?—24
What are the chances of a change in Division if an answer book is marked again ?—28
Six examiners compared on equivalent answer books—32
Distribution of marks—35
Reliability—42
Is the situation hopeless ?—48
Are examiners influenced by the number of pages written ?—49
The need for scaling—49

PART TWO—DETAILED REPORTS

5 *Review of the Literature*

Definition of terms used—55
Summary of early studies—56
Examination research abroad—58
Conversion of essay items to objective items—58
Reliability and validity of essay tests—60
Conclusions suggested by the foreign studies—64
Examination research in India—65
Functional aspects—65
Efficiency aspects—66
Variation in marking—67
(a) Variation of the same set of scripts—67
(b) Variations for equivalent sets of scripts—69
Misclassification of students due to marking errors—70
Reliability—71
(a) Content reliability—71
(b) Reader reliability—74
(c) Total reliability—75
Validity—77
Discussion—80
Suggestions for Improving Efficiency of Examinations in India—81

Improvement in curriculum—81
Improvement in questions—82
Specificity in questions—82
Increase in the number of questions—83
Reduction in the number of optional questions—83
Use of objective-type questions—85
Improvement in the accuracy of marking—86
Scaling of marks—87
Grace marks—88
Change in the basis of awarding "Passes"—88
Abolition of supplementary examinations—88
Internal assessment—89
Semester system—90
Viva-voce—90
Conclusions suggested by Indian Studies
Reform consciousness in India—91
Agencies active in examination reform—92
Publications pertinent to examination reform—92

6 *Data and Experimental Design* 96

Four Thousand Re Examined—96
Examination selected—96
Examiners selected—97
Experimental design—98
Selection of answer books—99
Preparation of answer books—100
Dividing and coding—102
Despatch and receipt—104
Card punching—104
Computations—105
Formulas used—106
Ninety Marking Ten—107
Choice of method of reproduction—108
Choice of examinations—108
Choice of answer books—109
Preparation of answer books—109
Experimental design—110
Despatch and receipt—111
Reliability of the data—111

7. *The Concept of "Reliability" as Applied to Traditional Examinations* 113
"Examiner", "content" and "marks" reliability—114
Examiner reliability—114
Content reliability—115
Marks reliability—115
"Correct mark" and "true mark"—115
"Standard error of marking", "of content" and "of marks"—116
Standard error of marking—116
Standard error of content—116
Standard error of marks—116
Effect of means and standard deviations—117
Standard errors "(scaled)" and "(unscaled)"—118
Mean errors "(scaled)" and "(unscaled)"—119
Range of marking errors "(unscaled)"—120
Applications—121
Determination and formulas—121
Examiner reliability—122
Content reliability—122
Marks reliability—124
Standard errors of marking and of marks (scaled)—125
Standard error of marking (unscaled)—125
Standard error of marks (unscaled)—126
Range of total error—126
Relationship to the studies in this book—126
A final note—127

8. *"Ninety Marking Ten"—Detailed Report* 128
Frequency distribution of marks—129
Classes or Divisions—132
Reliability of pass/fail—134
Means, standard deviations, and pass percentages of examiners—134
Frequency distribution of rank orders—139
Means, medians, and standard deviations—140
Correct mark—140
Standard deviations in Table 8.6—141
Standard error of marking—141
Average scale used by examiners—143

Total distribution—143
Interrelationships in Table 8 6 (b)—143
Reliability—144
Examiner reliability—145
Content reliability—148
Marks reliability—149
How well did the original examiners do ?—149
Factor analysis (1) answer books—151
Factor analysis (2) examiners—152
Position effects—154

9. *"Four Thousand Re-Examined '—Detailed Reports* 155

10 *Comparing the Four Examinations* 157

Mean marks—158
Standard deviations—160
Actual vs intended distributions—161
Note on use of medians for correlation coefficients—162
Correlation of number of pages vs marks—162
Mean of absolute differences—162
Examiner reliability—163
Content reliability—165
Marks reliability—167
Comparison with other studies—168
Standard error of measurement—170
Scatter charts for "Same" and "Different" re-examiner—171
Six examiners compared on equivalent answer books—171
Does the number of pages written influence marks?—175

11. *Detailed Statistics—History* 178

Distribution of marks—178
Table 11 2—179
Mean and standard deviation—179
Distribution of Divisions—180
Pass Percentage—181
Examiner reliability—181
Content reliability—182

Marks reliability—183
Standard error of measurement (SEM)—183
Differences—184
Scatter charts—186
Self-consistency of examiners' distributions—188
Item analysis—189

12. *Detailed Statistics—Hindi* 193

Distribution of marks—193
Table 12.2—195
Mean and standard deviation—195
Distribution of Divisions—197
Pass Percentages—197
Examiner reliability—198
Content reliability—198
Marks reliability—199
Standard error of measurement (SEM)—200
Differences—200
Scatter charts—202
Self-consistency of examiners' distributions—204
Item analysis—205

13. *Detailed Statistics—Biology* 208

Distribution of marks—208
Table 13.2—209
Mean and standard deviation—210
Distribution of Divisions—211
Pass Percentage—212
Examiner reliability—212
Content reliability—213
Marks reliability—213
Standard error of measurement (SEM)—214
Differences—214
Scatter charts—216
Self-consistency of examiners' distributions—218
Item analysis—220

14. *Detailed Statistics—Mathematics* 223

Distribution of marks—224

Table 14 2—225
Mean and standard deviation—226
Distribution of Divisions—227
Pass Percentage—228
Examiner reliability—229
Content reliability—229
Marks reliability—230
Standard error of measurement (SEM)—231
Differences—232
Reliability and scaling—234
Scatter charts—234
Self consistency of examiners' distributions—237
Item analysis—239

15 *Recommendations—Scaling* 242

What is scaling ?—243
Specified distributions of marks—244
' Raw" marks and ' scaled marks—245
The role of the Head Examiner—246
Randomization and equating—248
Randomization—248
Head Examiner re examines—248
All examiners examine a common batch—249
Use of a "link test —249
Scaling methods—250
Standard deviation scaling—251
Why not just add or subtract a constant ?—252
"Normalized ' scaling methods—253
The advantages of rank order scaling—255
Scaling of marks in different subjects to a common standard—258

16 *Recommendations—Question wise Multiple Marking* 260

Examiners who agree are likely to be wrong—260
How many examiners are needed?—262
Each examiner marks a different question—264

17 *Recommendations—General* 268

Six recommendations—269
In conclusion—270

APPENDICES

A *Content Reliability* 273

B *A Scientific Substitute for Grace Marks* 276

Relationship between "correct mark" and mark awarded by examiner—277
Determining the size of "measurement or marking error"—277
"Measurement error" in an examination—278
Setting the standard—278
What is likely to be the effect of using this system?—279
Summary—279
How to calculate passing probability"—280
Simplified calculations using Dr Taylor's Table—280
Notes—281
Supplement A — Distributions of error—282
Supplement B — Why probabilities are multiplied— 282

C *The Problem of the Crossed out Tick Marks* 285

D *Table of Reliabilities for* 80 *pairs of Examiners Marking* 50 *Answer Books Each* *opp.* 288

E *Computer Calculations Available* 289

F *Table of Marks and Ranks assigned by* 90 *Experienced Examiners to the Same Ten Answer Books in History* *opp* 290

BIBLIOGRAPHY 291

INDEX 305

Illustrations and Tables

(cross referenced)

Highlights of "Ninety Marking Ten"

Table 3 1 — Frequency distributions, totals, averages, and ranges of marks for each answer book (see Table 8 1 for details also *cf* Table 15 1 for effects of scaling) 9

Table 3 2 — Divisions awarded each answer book (see Table 8 2 for details) 11

Table 3 3 — Pass percentages of 90 experienced examiners 12

Table 3 4 — Frequency distributions of rounded ranks for each answer book (see Table 8 5 for details) 14

Highlights of "Four Thousand Re-examined

Table 4 1 — Frequency distribution of differences in marks between pairs of examiners 21

Table 4 2 — Frequency distribution of differences in Pass Percentages between pairs of examiners 23

Table 4 3 — Scatter charts of Divisions (see Table 10 2 for details) 26-27

Fig 4 1 — Can we trust Class or Division ? 29 30

Table 4 4 — Largest differences for six examiners of equivalent sets of answer books 32

Fig 4 2 — Equivalent sets marked by six examiners (see Table 10 3 for data) 33

Fig. 4 3 — Cumulative frequency curves (based on Table 4 5) 34

Fig. 4 4 — Graphs of frequency distributions (based on Table 4.5) 36

Table 4 5 — Frequency distributions for four papers (see Figs 4 3 and 4 4) 38 39

Fig. 4 5 — Graph of marks reliabilities (see Appendix D for data) 44
Fig. 4 6 — Graph of standard errors of measurement (see Appendix D for data) 45
Fig. 4 7 — Graphs of comparative marking errors 46

Other Research in India
Table 5 1 — Table of reliabilities obtained in Indian studies opp 70

Research Design
Table 6 1 — Experimental design opp 98

Reliability Concepts
Table 7 1 — Conceptual framework of reliability concepts opp 126

Details of "Ninety Marking Ten"
Fig 8 1 — Frequency distribution of sums of marks 130
Table 8 1 — Frequency distributions of marks for each answer book for each order of marking (summarized in Table 3 1) opp 130
Table 8 2 — Divisions awarded for each order of marking (summarized in Table 3 2) 133
Table 8 3 — Frequency distribution of means of marks 135
Fig 8 2 — Scatter chart of range vs mean 137
Table 8 4 — Frequency distribution of ranges and estimated standard deviations 138
Table 8 5 — Detailed distribution of ranks for both orders of marking (summarized in Table 3 4) opp 140
Table 8 6 — Means, medians, standard deviations, and ranks of marks and rank orders 142
Table 8 7 — Sample of inter examiner correlations 147
Table 8 8 — Correct mark vs Original mark 151

Comparing the Four Examinations
Fig 10 1 — How pooling data can produce a spurious correlation 169
Table 10 1 — Summary of the means and medians of various statistics (from Tables 11 2, 12 2, 13 2, 14 2) opp 170

Table 10 2 — Scatter charts of Division separately for 'Same and 'Different' examiners (summarized in Table 4 3) opp 172
Table 10 3 — Equivalent sets marked by six examiners (see Fig 4 2) opp 174
Table 10 4 — Correlations of marks with number of pages opp 176

History

Table 11 1 — Frequency distributions of marks (see also Fig 4 4) 178
Table 11 2 — Detailed statistics for all pairs of examiners (summarized in Table 10 1) opp 184
Table 11 3 — Scatter charts for examiners ABCD opp 186
Table 11 4 — Scatter charts for examiners EFGHIJ and X opp 188
Table 11 5 — Item analysis 191

Hindi

Table 12 1 — Frequency distributions of marks (see also Fig 4 4) opp 194
Table 12 2 — Detailed statistics for all pairs of examiners (summarized in Table 10 1) opp 200
Table 12 3 — Scatter charts for examiners ABCD opp 202
Table 12 4 — Scatter charts for examiners EFGHIJ and X opp 204
Table 12 5 — Item analysis 206

Biology

Table 13 1 — Frequency distributions of marks (see also Fig 4 4) opp 210
Table 13 2 — Detailed statistics for all pairs of examiners (summarized in Table 10 1) opp 214
Table 13 3 — Scatter charts for examiners ABCD opp 216
Table 13 4 — Scatter charts for examiners EFGHIJ and X opp 218
Table 13 5 — Item analysis opp 222

Mathematics

Table 14 1 — Frequency distributions of marks (see also Fig 4 4) opp 224
Table 14 2 — Detailed statistics for all pairs of examiners (summarized in Table 10 1) opp 232

Table 14 3 — Scatter charts for examiners ABCD opp 236
Table 14 4 — Scatter charts for examiners EFGHIJ and X cpp 238
Table 14 5 — Item analysis opp 240

Scaling
Table 15 1 — Effect of rank order scaling on marks in Table 3 1 257

Multiple Marking
Table 16 1 — Average marks of various numbers of examiners compared with correct mark 263
Table 16 2 — Tabulation of deviations of averages from correct mark 264
Table 16 3 — An improved multiple marking system (question wise marking) 266

Appendices
Appendix B— Table for grace marks opp 282
Appendix D— Reliabilities for 80 pairs of examiners (see Figs 4 5 and 4 6) opp 288
Appendix F— Actual marks assigned by each of the 90 examiners, and their ranks opp 290

PART ONE

Highlights

Highlight—moment or detail of vivid interest, outstanding feature

—*Pocket Oxford Dictionary*

1

Introduction

THIS BOOK reports on the reliability of examinations in India Under two different experimental conditions 4010 answer books in four different High School subjects were marked again by 130 examiners

Why do such studies?

There has been a tremendous amount of research on examinations in Europe and America Why waste time and money repeating this research in India? Specifically, why these particular studies?

Unfortunately, very few modern foreign studies apply directly to the Indian situation This is because most research minded examination systems have long since given up essay type examinations except for very specialised uses The research on which they based their decisions, most of it done decades ago, is no longer widely available Thus research on wide scale examinations, involving many examiners seems still to be of value

Also, somehow, research done elsewhere is never really convincing in a field like this There is a deep conviction among examiners that 'we do things differently Other examiners (especially examiners in other countries) may be unreliable, but *my* marking is always accurate. Even experts in educational measurement, fully aware of all the research results, may still suffer from this conviction that they, themselves, are somehow exempt from this charge of examiner unreliability

More basically, however, there is some reason to believe that the marking of answer books *should* be more reliable in India than abroad,

when done under similar conditions. In America, for example, education is locally controlled. There are (with one exception) no state-wide examination systems. Therefore, there are no generally-agreed-upon standards for the marking of papers—there is agreement neither as to what factors should be considered in marking, nor as to the general level of achievement expected. In India, on the other hand, the educational system within a particular state is uniform, with a fairly rigid curriculum. A state-wide examination system is built on the assumption that experienced examiners *do* have common standards. There is, therefore, some justification for traditionalists who reject Western examination studies saying, "These results do not apply to India—we do things differently here." Because of this, *Indian* studies are needed in India. Dr. H. J. Taylor, of Gauhati University, seems to be the only scholar who has responded in recent years to this vital need to study the *marking* of examination books under Indian conditions. (Dr. Gayen and others have studied marks, not marking.) Since Dr. Taylor's approach is somewhat different from ours, it is hoped that the two sets of studies can be considered complementary to each other.

One of the two studies, in which 90 experienced examiners re-marked the same ten history answer books, has a special justification. One criticism of the only western studies of this type of which we were aware when we began this research, is that they examined the re-marking of only a single answer script, in each subject. Everyone "knows" that examiners differ in standard. Might it be, however, that a group of examiners could at least agree on which of a *pair* of answer books is better—or, more generally, agree in placing a group of answer books in the same order of merit? If there is such agreement, then the disagreement on actual marks is irrelevant, as these can be adjusted by "scaling". Thus it is important to know if agreement on order of merit exists.

This study shows that unfortunately even scaling would not change the fact that traditional essay-type examinations are *unreliable.* This unreliability has been proven again and again. Yet we continue to be convinced that each one of us, personally at least, can mark an answer book accurately. Why?

It is interesting to speculate about the sources of this conviction of personal perfection in examination marking. Is it because we are unable to face the consequences of *not* believing that our marks are reliable? Is it because, after the tremendous investment of time and effort in marking a set of answer books, we *must* believe that we have marked them correctly—because to admit that all that time and effort have really been wasted is almost intolerable? Perhaps in most areas in which personal judgements are demanded, especially those judgements which affect the lives of others, we are almost forced psychologically to a conviction of

infallibility This conviction is the only defence against an overwhelming sense of personal responsibility, if we admit that our judgements may be seriously wrong History is filled with instances of wrong judgements by persons in high authority How many have ever publicly admitted that they were in error?

Ultimately, perhaps every skeptical examiner may need to be forced to re mark his own papers before he will *really* believe that he himself is unreliable Fortunately, many who are less self important will realize that the results shown in this book apply equally to their own personal work This is the justification for the present studies

2

Briefly Summarized

THIS BOOK reports two different studies Both are based on the actual answer books written by Class X students in Higher Secondary Board Examinations in India Both involve experienced examiners who have worked for several years under the same Higher Secondary Board

One is the "ninety marking ten" experiment Ninety photographically exact copies were made of each of ten history answer books So each of the ninety experienced examiners marked the same ten answer books They marked them as carefully as possible

The other study is called the 'four thousand re-examined" study Four thousand answer books were selected—one thousand each in history, Hindi, biology, and mathematics (geometry) Forty examiners—ten in each subject—were asked to mark them a second time Each examiner marked 50 answer books that he himself had previously marked, and 50 that had been marked previously by another examiner The answer books were mixed up, so the examiner did not know which were his and which were another's

Now for some of the findings

What happens when ninety experienced examiners mark the same ten history answer books?

** One of the answer books was considered the *best* of the ten by one experienced examiner, and the *worst* of the ten by another

** The only answer book that was considered worthy of a *Distinction* by one experienced examiner, was *Failed* by seven Eight gave

it a First Class Marks ranged from 22% to 76% So whether a candidate Fails or gets a Distinction may often be as much due to luck (i e which examiner marks his paper) as to ' merit

** There was considerable variability in the range of marks assigned each of the ten answer books The *narrowest* range was 13 marks out of 50 (26%) between the highest and the lowest mark assigned to the same answer book by the 90 experienced examiners The *widest* range was 28 more than twice as much

** The 90 experienced examiners did not agree on the average merit of the 10 candidates as a batch Their standards differed widely The *average* marks awarded the group ranged from 8 to 22 out of 50 i e from 16% to 44% Pass percentages ranged from 10% to 80%

** The number of First Classes awarded by some examiners was larger than the number Passed by others

** Eighty six per cent of the examiners agreed as to which was the best of the ten answer books However they disagreed widely on how many marks to award this best answer book The marks awarded this candidate ranged from 17 to 35 out of 50 i e from barely passing to high First Division

** There was much less agreement as to which were the poorest candidates

** There was wide disagreement even on which Class or Division should be awarded each candidate Several of the candidates were awarded all three Divisions The highest number of examiners awarding a candidate the same Division was only 50 or a little over half of the 90 examiners Fortunately, however, the aggregate total Division on any examination is probably more reliable than the Division on any single paper

** For only two out of the ten did all examiners agree on whether they should pass or fail This means that for 80% of the candidates there was disagreement as to whether they should be passed or failed And the higher reliability of the aggregate does not help here at all since failure on a single paper can mean failure in the examination

** The average of the 90 marks was considered the fairest mark for each candidate So marks which were different from the average were considered errors The range of errors, for the 90 experienced examiners on 10 answer books was from zero error on one answer book (but never zero on more than one) for 12% of the examiners to as high as 15 (which is 30% marks) The *average* error when only one examiner marks a batch of history answer books is 2 9 marks out of fifty Suppose we want to guarantee that the error will be less than two marks (out of 50) for at least two-thirds of the candidates Each answer book would have to be marked by *five* examiners and their results averaged

** Probably 4,000 out of every lakh of candidates are awarded marks which are 15% or more above or below what they should have received—i.e. they are one full Division lower or higher than the marks they would have received if their answer books were marked without error.

What happens when forty experienced examiners mark four thousand history, Hindi, biology and mathematics (geometry) answer books a second time?

** If four thousand answer books are marked a second time, about 52 of them will be raised or lowered by two Divisions. For example, Firsts will receive a Third, and Failures will be awarded a Second. Even in mathematics, 11 out of one thousand will be changed by two Divisions. If these candidates are administered a second examination, the number of two-Division changes will be much larger.

** On the average, examiners who mark the same candidates twice, will give the same Division or Class both times to only about 65% of the answer books. For thirteen out of one thousand, the two marks will differ by two Divisions.

** Mathematics examinations are only slightly more reliable than history examinations. The agreement on Divisions is 77% for geometry and 60% for history. Objective-type examinations, on the other hand, give an agreement above 90% for two *separate* examinations—and, of course, 100% for two examiners of the same answer books.

** Examiners are inconsistent in pass percentage, the number of candidates awarded each Class or Division, and even the marks awarded individual candidates. They not only disagree with each other, but they even disagree with themselves when re-doing their own work.

** Out of 229 First Divisioners, 83 were awarded a Second or Third by another examiner. Twenty-two (mostly in geometry) were raised to a Distinction.

** Twenty-seven per cent of all candidates *failed* by one examiner were *passed* by another examiner.

** The average difference between examiners is nearly six per cent of marks. Some pairs of examiners—even in geometry—differ by as many as 20 marks out of 50.

** Examiners changed the pass percentage by as much as 58% when they re-marked a batch of answer books. Even a single examiner's pass percentage varied as much as 58% from one marking to the next.

** When an examiner marks answer books a second time—and he marks some that he himself had marked first, and some that were first marked by another examiner—he will agree with his own first marking only slightly better than he will agree with another's marking. Examiners in history, Hindi, biology and geometry all agreed with each other only

slightly less than they agreed with themselves But in all cases, the agreement was low

** The Division awarded a candidate depends as much—and sometimes even more—on chance (i e who his examiner is) as on his own merit (This applies to any single paper A *part* of these chance factors are averaged out, when marks in several papers are added together However, if a candidate *fails* "by chance" he fails the entire examination Grace marks do not really compensate for the effects of examiner unreliability)

** Answer books can be reliably differentiated into only about five or six categories Elimination of the "100 marks" scale would tend to increase reliability of marking

** Allowing a choice of questions lowers the reliability of an examination Paper setters apparently cannot judge the difficulty of most questions accurately Sometimes the marks of very good candidates are lowered because they like to choose hard questions while the marks of weak candidates are raised by a lucky choice of easy questions

** In spite of popular belief to the contrary, examiners are *not* influenced by the number of pages written

** Different scales seem to be used for different subjects Just as one degree Farenheit is not equal to one degree Centigrade so one mark in Mathematics is not equal to one mark in Hindi Marks in different subjects should not be added together, unless they are changed or converted into the same scale

** In comparison with results from other countries (France, England, U S A), Indian examinations seem to be slightly *more* reliably marked than those abroad This statement applies only to untrained examiners Studies abroad indicate that special training can greatly increase the reliability of marking Thus it is not the examiners, as such, that are unreliable It is this particular system of examination (essay-type), the lack of properly written questions the lack of proper scaling, and the lack of adequate instructions and training for examiners which are the sources of the unreliability of marking revealed by these studies

3

Highlights of "Ninety Marking Ten"

THIS STUDY set out to answer the question, "What happens when one hundred experienced examiners are asked to mark the same ten answer books ?"

Ten history answer books were selected. The original examiners' marks of these ten answer books spread from 7 to 28 marks out of 50. Forty per cent had passed and sixty per cent had failed, according to the original examiners.

We especially selected answer books in which the candidates had answered the *same* five (out of ten) questions. This was in order to reduce the variability due to different choices of questions. *This makes the reliability of the examination in this study higher than the regular public examination in history.* (The reliability when different candidates choose different questions is reported in the "four thousand re-marked" study.)

The instructions issued to the examiners were the same ones which had been issued for the regular examination. It was not felt necessary to send any "model answer books", as all these examiners had already marked the same examination several months earlier.

Ten examiners did not return their marks sheets, so this study is based on the remaining ninety. All ninety examiners were regular experienced examiners of the same Board of Secondary Education in India. They marked the examination books as carefully as possible.

1. *What happens when ninety experienced examiners mark the same ten history answer books ?*

Table 3.1 shows the marks awarded to each of the ten answer books. These are marks out of 50.

TABLE 3 1

FREQUENCY DISTRIBUTIONS OF MARKS AWARDED BY 90 EXPERIENCED EXAMINERS TO EACH OF 10 HISTORY ANSWER BOOKS

	Answer Books										
Marks	D	E	C	I	F	A	B	H	G	J	Total
38		1									1
37		—									—
36		—									—
35	1	—									1
34	5	—									5
33	4	—									4
32	7	1									8
31	6	4									10
30	12	3	1								16
29	2	4	1								7
28	5	2	—								7
27	7	4	1								12
26	6	8	1	1	1						17
25	9	10	2	1	1						23
24	4	4	3	—	—						11
23	6	9	7	—	—	1					23
22	3	4	11	3	2	1					24
21	3	9	8	1	6	2					29
20	4	10	9	6	1	2					32
19	3	3	10	2	9	1		1			29
18	2	2	8	7	4	12		2			37
17	1	5	4	18	15	12	2	4	2		63
16		3	4	3	4	5	3	3	1		26
15		2	5	14	10	7	4	7	2	3	54
14		—	4	11	8	12	8	3	2	4	59
13		1	3	2	10	9	12	7	5	5	54
12		—	4	11	3	6	14	9	13	6	66
11		1	2	4	5	7	12	17	13	8	64
10			1	?	4	8	11	9	8	10	53
9			—	1	3	2	10	14	14	8	52
8			1	1	2	1	5	7	9	14	40
7				2	1	2	8	4	13	13	43
6					1		—	1	3	7	12
5							1	3	2	6	12
4								3	2	1	6
3								1	—	3	4
2									1	2	3
Total	90	90	90	90	90	90	90	90	90	90	900
Average	27 0	23 1	19 0	15.5	15 4	14.5	11.2	10 9	9.8	8 8	15.5
Range	19	28	23	20	21	17	13	17	16	14	

For simplicity, the ten answer books (A to J) are arranged in order of their mean (average) marks

For example, the first column shows the marks awarded by 90 experienced examiners to candidate D One examiner awarded him 35 marks Five examiners awarded him 34 marks At the bottom of the column, we find that one examiner awarded him 17 marks out of 50, two awarded him 18, etc At the very bottom we see that the average mark awarded by the 90 experienced examiners to answer book D was 27 marks out of 50 And the range of marks, from the lowest to the highest, is 19

The lines across represent the Classes or Divisions Those below the first line have failing marks Those between the first and second lines are in the Third Division The top line, above 38 marks, is a Distinction

Look at the second column, for answer book E This candidate was awarded 38 marks out of 50 by one examiner—and only 11 marks out of 50 by another The only one of the ten candidates who was thought worthy of a Distinction by one examiner was failed by seven

The only candidate who was failed by all the examiners was Candidate J, and even he was awarded a spread of marks from 2 to 15

The range of marks is interesting There was such wide disagreement that some answer books received marks over a range of 28 out of 50 Even where there was the maximum degree of agreement among examiners (candidate B) the range was still 13 marks out of 50

This range is somewhat related to average marks Those with the highest averages were awarded the largest range of marks However, look at candidates B and H Their averages were both 11 marks out of 50, but their ranges were 13 and 17 There are some qualities on the importance of which examiners agree, and some on which they disagree When an answer book contains mainly the former qualities, the range of marks will be narrower

What does all this mean? Teachers and students are rightly concerned with the maximum error possible in the present system of examinations Answer book E shows this (If we had had 100 answer books instead of only ten, the maximum error would probably have been even larger) Obviously, the correct mark for this candidate is 23, the average of the 90 experienced examiners Suppose that in a large public examination there are, say, 1,500 candidates who, like E, "should' receive 23 marks out of 50 Out of these, as many as 17 may receive a Distinction, while about 117 of them will fail—even though they are all actually of *equal* merit

2 Are these differences just random errors ?

Of course, there is the possibility that examiners made just one mistake out of ten It is possible that they agreed on almost all candidates,

but just happened to make a mistake on one—and that if this one mistake was for a different candidate for each examiner this could produce the wide variations in Table 3 1

Is this what happened ?

No There are statistical proofs of this But if you want to see for yourself you can look at Appendix F where all the marks of all the examiners are listed

Look at this pair Here are two experienced examiners (Nos 64 and 69) who though they awarded almost exactly the same average marks differed widely on many of the candidates

Candidate	First examiner	Second examiner	Difference
E	38	20	18
D	29	20	9
I	26	15	11
C	16	20	4
H	12	15	3
F	10	19	9
A	10	18	8
B	9	13	4
G	9	13	4
J	7	13	6
Mean	16 6	16 9	7 9

3 *Can we trust Class or Division ?*

The answer to this question is already obvious from Table 3 1 But it is sharpened up in Table 3 2

TABLE 3.2

FREQUENCIES WITH WHICH EACH CLASS OR DIVISION WAS AWARDED BY 90 EXPERIENCED EXAMINERS TO EACH OF 10 HISTORY ANSWER BOOKS

Class or Division	Answer Books D	E	C	I	F	A	B	H	G	J	Total
Distinction I		1									1
II	35	8	1	2	2	1					44
III	39	41	15	37	37	30	2	7	2		100
Fail	16	33	50	47	44	54	64	50	44	36	214
10-16 marks		7	23								369
0-9 marks			1	4	7	5	24	33	44	54	172
Total	90	90	90	90	90	90	90	90	90	90	900

Each row in the table represents a single Division or Class. Each Division includes 6-8 marks, but Fail includes twice as many Therefore we have divided Fail into two groups, 0 9 and 10-16 marks

Candidates E and C were each placed in five different categories, by different examiners All others were placed in at least three categories, except J Candidate J was so weak in the subject (average marks 8 8 out of 50) that he was failed by all examiners

The highest number of examiners awarding the same Division to a candidate was 50 out of 90, which is only a little over half Obviously, statements such as, ' He is a First Divisioner" or ' He is a Failed student" do not carry a very accurate meaning As we shall also see in the second

TABLE 3 3

VARIATION IN PASS PERCENTAGES WHEN 90 EXPERIENCED EXAMINERS MARK THE SAME 10 HISTORY ANSWER BOOKS

1	examiner	passed	8	candidates	(80%)
4	examiners	,,	7	,	(70%)
14	,,	,,	6	,,	(60%)
16	,,	,	5	,,	(50%)
15	,,	,	4	,,	(40%)
23	,	,,	3	,,	(30%)
13	,	,,	2	,,	(20%)
4	,,	,,	1	candidate	(10%)

study, Class or Division depends nearly as much on pure chance, (that is on which examiner marks the candidate's answer book), as it depends on the candidate's own actual merit

4 Can we trust "Pass Percentage" ?

We often evaluate teachers, and educational institutions, in terms of their pass percentages Is this fair ?

Remember that under the present system, the history answer books of one higher secondary school will go to one examiner, while those of another will go to a different examiner (The only exception is Assam where, in some examinations, the answer books are mixed up together before sending out) Under these circumstances, what does the "pass percentage" really tell us about a teacher or institution ?

Table 3 3 shows how widely experienced examiners will differ on the pass percentage of the same batch of answer books For one examiner the pass percentage was 80%, while for another four it was only 10% For 19 examiners it was above 60%, while for almost the same number (17) it was below 20% Obviously, *pass percentage tells us more about the examiners than it does about the candidates or their class teachers*

The variation in pass percentage from year to year in the public

examinations, and the variation from institution to institution and from teacher to teacher, just tells us that the examinations are unreliable Taylor's (1962e, 1963a) studies have shown similar results Readers may know that Dr Taylor, who was then Vice Chancellor of Gauhati University, was formerly a professor of physics (not of psychology or education) He approached the problem of examinations as measure ments in the same rigorous way as he would approach the problem of the reliability of any measurement in physics His conclusion about the pass percentage (1963, page 23) was

> "Whether they encounter high pass or low pass examiners is a matter of chance, and on that their fate depends This overwhelming element of chance, the extent of which has not previously been appreciated, is perhaps one of the main causes for the failure rate of 50% to 70% in most of the large public and University examinations in India"

Dr Taylor's data indicate that reducing this chance element by "randomization and scaling" (a scientific method which will be explained later) also reduces the failure rate by as much as one quarter This clearly implies that *one-quarter of all failures are by chance alone*

5 *Can examiners agree on who is best and who is worst?*

Of course "everybody knows' that examiners do differ in standard (This is what we hear *after* we have proved that it is true) Even though these were experienced examiners, of the same Board, working under the same instructions of the same Head Examiner, their standards still differed quite widely

However, if examiners can at least agree on the *relative* merit of a series of answer books, then the situation is not serious If we know which is best and which is worst, we can award them Distinction and Zero, and award the rest marks in between (See chapter on Scaling for a more detailed discussion of the methods which can be used)

What we did was this We looked at each examiner s marks sheet separately The highest mark awarded by the first examiner was given a rank of "1". The next mark of this same examiner was ranked "2", and so on down to "10' for the first examiner's marks Then we did the same for the second examiner, and the third, on through the ninety

Table 3 4 shows the results Answer book D was considered best (rank 1) by 77 of the 90 experienced examiners Twelve considered it the second-best of the ten, and one considered it only third best At the bottom, you can see that the average rank (or order of merit) of candidate D's marks was 1 2

For convenience, the answer books in Table 3 4 are in the same

order as Table 3 1 The agreement on ranks is very different from the agreement on marks

The highest degree of agreement in ranks was for candidate D Eighty-six per cent of the examiners considered him the best of the ten (Actually, in seven of these 77 marks sheets, this candidate tied with another for first place See Table 8 5) But look at Table 3 1 Even if the examiners agreed that D was best, they certainly did not agree on how many marks to award him Thirty-five awarded him a First, 39 a Second, 16 a Third, and his marks ranged from 17 to 35 out of fifty

The least agreement was for candidate F Some experienced examiners considered this the best of the 10 carefully marked answer books, and some considered it the worst F's marks, however, did not range as widely as some others

TABLE 3 4

FREQUENCY DISTRIBUTION OF THE RANKS (ORDERS OF MERIT) AWARDED BY 90 EXPERIENCED EXAMINERS TO EACH OF 10 HISTORY ANSWER BOOKS

Rank Order	Answer Books									
	D	E	C	I	F	A	B	H	G	J
1	77	16	2		1					
2	12	59	14	3	1					
3	1	12	49	13	11	4		1		
4		2	18	28	32	25	1	—		
5		1	5	19	24	25	3	5	1	
6			1	19	13	25	11	13	6	1
7			1	4	5	7	38	25	11	10
8				4	2	4	19	22	26	10
9					—		10	18	24	26
10					1		8	6	22	43
Total	90	90	90	90	90	90	90	90	90	90
Average Rank	1 2	2 1	3 3	4 9	4.9	5 4	7.7	7.8	8 6	9.2
Range	3	5	7	7	10	6	7	8	6	5

NOTE The averages are from Table 8 5 where ties are listed as half-ranks.

At the bottom of Table 3 4 you will find the range of ranks awarded each Notice that they vary from 3 to 10 ranks, with an average of 6 4 ranks out of ten Thus the highest degree of agreement (D) still covered nearly one third of the range of ranks, while the average included nearly two-thirds of the available ranks

Again, compare Table 3 4 with Table 3 1 In Table 3 1, the highest agreement on marks was for the weakest candidates But in Table 3 4, the

highest agreement on ranks is for the best candidates Is it more important to be sure that you fail all the weak students or is it more important to be sure that the best students are awarded the marks that they deserve ?

This is, of course, a question of educational philosophy that is beyond the scope of this book However, it can be argued that 'the greatest good of the greatest number is more dependent on the leadership of those at the top of the intellectual range, than on those of only average ability If so, then it is rank orders and not marks that we should request from examiners The Board can then assign marks to these rank orders, according to accepted scientific procedures (see chapter on Scaling) Under such a system, Candidate D s marks would not have ranged from 17 to 35 They would have ranged only from 20 to 30 (or, if we dropped the single third rank, from 23 to 30), and the average mark would have been 29 (instead of 27)

6 *What element does "chance" play in passing and failing ?*

We have already quoted Taylor's statement on "the overwhelming element of chance", and the results discussed so far support this But perhaps it needs to be spelled out more specifically, in relation to the pass fail line

The most crucial decision for an individual (as against the crucial decision for society which we have discussed above) is whether the individual passes or fails And failing on a single paper can of course, mean failure on the examination

Back, again, to Table 3 1 It is reasonable to assume that the average mark awarded by 90 experienced examiners is the 'correct" mark for a candidate

The lowest "pass ' was candidate C, whose ' correct mark' is 19 0 But with 38 per cent of marks, he does clearly deserve to pass Out of 90 examiners, 24 (or 27% of the examiners) would have failed this candidate if ' by chance" they had been his only examiners

Candidates I and F, with "correct marks' of 15 5 and 15 4 clearly deserve to fail Yet 78 of the 180 marks awarded by the 90 examiners to these two candidates would have passed them Thus 43% of the examiners passed and only 55% of them failed these two candidates who, so far as our evidence goes, clearly deserved to fail

Perhaps passing a candidate who ought to fail is less serious than failing one who ought to pass In a regular Board examination, candidate C's answer book would have been assigned to only one single examiner Thus, though his true merit is 38 per cent of marks, there was one chance in four that he would have been failed *only* because ' by

chance" his answer book was assigned to a particular examiner (If his examiners reported only rank orders, however, and these were scaled, he would have been failed by only 7 instead of 24 examiners)

There is, of course, a degree of unreliability in any examination The point we are trying to make here is that this unreliability, this "chance factor, can be greatly reduced The entire science of educational measurement is at the disposal of the Indian examination system Is it not time that we begin to use it ?

7 How reliable is this examination ?

In the layman's sense of the term, we have been discussing reliability (and unreliability) all along But many readers may be used to the more technical use of the term "reliability" as it relates to tests and examinations Particularly, they are interested in two specific measures of reliability, the coefficient of reliability and the standard error of measurement (The non statistical reader will find a layman's explanation of the standard error of measurement in the next chapter, where it is more crucial to the discussion)

We must distinguish between the reliability with which an examiner marks an examination, and the reliability of the examination itself (See Chapter VII)

The average "examiner reliability" or "reliability of marking" of this history examination was 83 In comparison with this, the "reliability of marking' of an objective examination is always 1 00 For the inability of two examiners to agree on the marks they assign is not the only source of unreliability in an examination There are also factors related to the examination questions themselves Thus the reliability of the actual marks assigned students is due to a combination of these factors We have called this reliability (which is comparable to the usual use of the term "reliability" in objective examinations), "marks reliability" (See Chapter VII for fuller discussion)

The marks reliability of this examination is 71 The technically trained reader will recognize that this is quite inadequate for the purposes for which examinations are given, i e to distinguish reliability between individuals

This is underscored by the standard error of measurement, which is 3 43 This means that 5 out of every 100 candidates will be awarded marks which are either about 7 marks above or 7 marks below what they would have received on a perfectly reliable examination

Unfortunately, all of these reliability figures are misleading, because they are based on the assumption that all examiners are using the same "scale" in marking their examinations—in the layman's terms,

that they are all 'marking to the same standard But they are not For example, suppose a psychological test has a coefficient of reliability of 83 and a standard deviation of 5 6 (the average of Examiners 15 and 6) What would we conclude ? In these circumstances the standard error of the difference—which tells us the difference between two sets of marks—is 3 26 marks out of 50 Thus the average difference between two examiners should be 2 1 marks and the maximum about 8 marks out of 50

Now let us look at an actual pair of examiners whose examiner reliability is 83

Candidate	Marks by Examiner		Difference
	15	6	
A	7	23	16
B	5	15	10
C	8	27	19
D	20	34	14
E	11	29	18
F	6	26	20
G	2	16	14
H	7	17	10
I	10	22	12
J	5	13	8
Average difference			14 1

Notice that Examiner 15 passed only one candidate while Examiner 6 passed seven The average difference between the two is 14 marks out of 50 (not 2 1) and the largest difference is 20 (not 8) This is the penalty for not using modern methods of scaling

It must be remembered also that these figures (i e the examiner reliability of 83 and marks reliability of 71) were obtained for a set of answer books in which all candidates had answered the same five questions Had the usual situation prevailed—different candidates answering different combinations of five out of nine questions—the correlations between examiners would have been lower The highest marks reliability that we found for history where a choice of questions was allowed was 43 as against 71 when all candidates had answered the same five questions (See Figure 4 5)

8 Will it help to have several examiners mark each answer book and average the results ?

Multiple marking of answer books is the only way to reduce the unreliability of marking to the minimum And this also increases total reliability, as examiner unreliability is one of the factors in the total situation

To study what would happen, we assigned examiners "at random" to candidates For example for Candidate A we pulled out the mark-sheets of 10 of the 90 examiners We wrote down, in the order in which the marks sheets had been drawn, the marks assigned by each to candidate A Then we put these back, and again drew another 10 examiners for candidate B Thus we obtained a series of ten marks for each candidate

Then we averaged the first two marks, then the first three marks, then the first four marks etc for each candidate We compared these averages with the 'correct mark based on the average of all 90 examiners

How many examiners do we need? This depends partly on how accurate we want to be Suppose we say that the "error of marking should not be more than 2 marks (out of 50), for at least two thirds of the candidates Our results indicate that each answer book would have to be marked by *five* different examiners to reach even this modest standard of accuracy

4

Highlights of "Four Thousand Re-examined"

FOUR THOUSAND Class X final Examination answer books were marked a second time

One thousand of these came from each of the following subjects history, Hindi biology mathematics

Half of the answer books were marked a second time by the same examiner—the other half of them were marked the second time by a different examiner

In each subject all examiners had worked for the same Board and in fact for the same Deputy Head Examiner

The instructions issued to the examiners were the same ones which had been issued for the regular examination It was not felt necessary to send any model answer books as all these examiners had already marked the same examination several months earlier

1 Differences between examiners

The differences between the two markings of each of 4 000 answer books ranged from 0 to 20 marks out of 50 Surprisingly it was a mathematics (geometry paper) examiner who differed from his own first marking by 20 points Examiners re-examining their own answer books did only *slightly* better than examiners re examining some other examiner s answer books The *average* difference between pairs of examiners of the same answer books were as follows

History 3 0 out of 50 marks

Hindi	2 8 out of 50 marks
Biology	2 8 out of 50 marks
Mathematics	2 6 out of 50 marks

It is interesting to note that mathematics, an "exact" subject, shows almost as much average difference between examiners as biology, which depends more on judgement As will be pointed out later, this is due to differences in the scales used for these two subjects

These average differences may look small But one does not expect to be able to climb Mount Everest just because the *average* height of the Himalayas may be only 10,000 feet In other words, each of these averages represents a wide range of differences between examiners (See next section) And the 'difference" for any particular candidate can be disastrously large

This is shown by the fact that examiners frequently fail to award even the same class or division to a candidate in both markings Sometimes a candidate who was awarded a Second Class by one examiner was failed by the other—sometimes a Third Divisioner was raised to a First (See Table 4 3) Overall,

History	40 1%
Hindi	32 0%
Biology	38 5%
Mathematics	22 9%

were awarded different Classes or Divisions when they were marked a second time

2 How large a mistake is likely to be made for any single candidate?

Table 4 1 shows how much examiners differ from each other, when marking the same candidate It is a frequency tabulation of the number of differences of each size appearing when (a) the same examiner re-examined his own answer books, and (b) the answer books were re-examined by a second examiner In each case, the frequency distribution is based on ten pairs of 50 answer books each

For example, look at the history columns When pairs of examiners (the D column means "different') marked the same answer books, in one case they differed by 16 marks out of 50 In two cases they differed by 12 marks They agreed completely (0 marks difference) in only 46 cases out of 500

Now look at the mathematics (geometry) column Discouraging, isn't it, to discover that even in mathematics an examiner can disagree with himself by as many as 20 marks out of 50, i e 40% of marks! Granted that this is only one case out of 500 But the fact that this is not just a single random, or freak error is shown by the fact that there are several other errors nearly as big—18, 17, and 16 marks This means that if there are 50,000 candidates, we can expect that as many as 75 to 100 of

TABLE 4 1

Frequency Distributions of the differences in Marks Awarded by the First and Second Examiners of Four Thousand Answer Books

Examination		History		Hindi*			Biology		Maths		History	Hindi	Biology	Maths
Re-examined by Same or Diff examiner		S	D	S		D	S	D	S	D	S+D as accumulated totals			
No of answer books		500	500	500		500	500	500	500	500	1000	1000	1000	1000
	20								1					1
	19													
	18								1					2
	17								1					3
	16		1						1		1			4
	15		1							1	3			5
	14									2				7
	13			(9)		1		1		2		1		9
	12		2	(8)	2	5	1			2	5	8	2	11
Size of difference	11		6				1	1	3	4	11		4	18
(marks out of 50)	10	3	9	(7)	4	6	2	4	2	3	23	18	10	23
	9		14	(6)	3	12	5	9	5	7	37	33	24	35
	8	8	17				7	11	7	19	62		42	61
	7	9	22	(5)	11	21	12	19	6	23	93	65	73	90
	6	18	29	(4)	22	46	19	24	7	29	140	133	116	126
	5	39	53				39	49	19	39	232		204	184
	4	58	72	(3)	76	70	79	52	36	39	362	279	335	259
	3	84	71	(2)	108	118	75	78	59	68	516	505	488	386
	2	99	64				90	88	82	79	679		666	547
	1	106	93	(1)	166	145	104	102	148	112	878	816	872	807
	0	76	46	(0)	108	76	66	62	122	71	1000	1000	1000	1000

* Hindi was marked out of 34 Each mark was multiplied by 50/34 to make it equivalent to the rest of the table The original marks are put in parentheses in line with their rounded equivalents For example 50 34×8 marks equals 11 8 marks equals 12 marks when rounded

them can be victims of errors close to 40% of marks

Table 4 1 speaks for itself, and the experienced teacher (or examiner) will be able to extract a great deal of interesting information from a detailed study of it And Table 3 1, with many more examiners involved, showed that even larger errors (up to 27 marks out of 50) are occasionally possible

Fortunately, the very extreme variations are not very frequent The average differences (as pointed out earlier) are not too large Still, "Many persons are drowned in rivers whose average depth is only two feet" It seems to us that we are rightly much more concerned with the extreme cases than with the averages *If even one per cent of higher secondary school examinees can have their entire future ruined by these large random errors, can we justify this by saying that the "average" error is not more than 5 or 6 marks?* Such large random errors just do not occur when modern, scientifically based examination methods are used

3 Can we trust "Pass Percentage"?

A great deal of weight is placed on the 'pass percentage" Teachers are condemned if the pass percentages in their classes fall, praised if they rise Institutions are compared by their pass percentages, and nothing else This is, to say the least, incredibly naive There should be at least some attempt to take account of the quality of the students admitted, before deciding that the pass percentage tells you how good a teacher or institution is It is well known, in fact, that some institutions raise their pass percentages just by refusing to admit any student who has any possibility at all of failing

Aside from this, is the "pass percentage" actually a reliable figure? It would be reliable only if it could be shown that it is a constant standard —i e that examiners agree, with themselves and with others, on the pass percentage

We saw in the last chapter that ninety experienced history examiners varied widely in pass percentage for the same ten answer books But what about this study, where ten examiners are paired with each other, and with their own previous records? And what about subjects other than history? Can we find any reason, here, to discount the results of the previous study?

No Table 4 2 makes this quite clear Table 4 2 shows the *difference* in the percentage of the 50 answer books passed, in the first and second marking In Hindi, for example, there was one *difference* of 58% between the percentage passed the first and second time by the *same* (S) examiner Examiner D had passed 66% of his candidates the first time he marked them The second time he marked this same batch of 50 answer books,

he passed only 8% of them This difference (66%–8%=58%) is the difference tabulated in Table 4 2

The highest difference for Hindi when the second marking was by a different examiner (D) was 54% In fact, in 20 batches of 50 history answer books each, only two batches had the same pass percentage the first and second times they were marked

The situation was sightly better in history than in Hindi, and consi derably better in biology But even in mathematics, one examiner differed as much as 12% with himself, in pass percentage on two markings of the

TABLE 4.2

FREQUENCY DISTRIBUTIONS OF DIFFERENCES IN 'PASS PERCENTAGE' BETWEEN FIRST AND SECOND EXAMINERS OF EIGHTY SETS OF FIFTY ANSWER BOOKS EACH

Subject	History		Hindi		Biology		Maths		Total	Accumulated total
Re-marking by Same or Diff examiner	S	D	S	D	S	D	S	D		
No of sets of 50 answer books	10	10	10	10	10	10	10	10	80	
Difference in Pass Percentage" between first and second marking 58%			1						1	1
56%										
54%				1					1	2
52%										
50%										
48%										
46%		1							1	3
44%										
42%										
40%		1							1	4
38%				1					1	5
36%	1								1	6
34%										
32%	1	1							2	8
30%	1								1	9
28%			1						1	10
26%		1							1	11
24%						1			1	12
22%			1			1			2	14
20%										
18%				2	1				3	17
16%			2						2	19
14%		1	1			1			3	22
12%	4	1					1		6	28
10%			1		2	1		1	5	33
8%		1		1				1	3	36
6%	1			2	3	1	2	1	10	46
4%				1	1	2	4	4	12	58
2%	2	2	1	2	3	2	1	1	14	72
0%		1	2			1	2	2	8	80

For example, If one examiner passes 76% of his 50 candidates, and the other examiner passes 60% of the same candidates the 'Difference in 'Pass Percentage'" is 16%

same answer books And only four out of 20 pairs, even in geometry, had the same pass percentage the first and second times.

In the "Total" column we can see that *only 8 out of 80 pairs had the same pass percentage the first and second time.*

The last column accumulates the totals. Thus we can see that 19 out of the 80—i e. one-quarter—differed in pass percentage by 16% or more. Nearly half of them differed by 8% or more. Of course, with large groups, the differences tend somewhat to "average out". Still, it is evident that *not much reliance can be placed on pass percentage.*

4. Can we trust Class or Division ?

Errors in exact marks would not be so serious if we could at least trust the broad Classes or Divisions If a "First Divisioner" is *really* a First Divisioner, then there may be some predictability in the academic world.

Several studies done in India have correlated subject marks at various successive stages of education, longitudinally. In general, the correlasions have been low, indicating only a minimum of stability over time. The present study approaches the problem from the angle of having the same answer books marked twice.

In the previous chapter we saw that when a large number of examiners mark the same answer book, there is a wide spread of Divisions. In some ways it is perhaps more realistic to consider only a pair of examiners, since double-marking actually could be carried out routinely.

Suppose a candidate is awarded a particular Division. What Division is he likely to be awarded if the same, or a different, examiner marks his answer book a second time ?

Table 4.3 answers this question. For this table, we have pooled the results of both the 500 answer books re-examined by the original examiner, and the 500 re-examined by a different examiner. (These will be reported separately later.) Pooled psychological data, particularly correlational data, are frequently suspect. In this case, however, the pooling is clearly justified. Examiner A's Third Divisioners may be awarded a Second Class by Examiner B, while Examiner C's First may be awarded a Third by Examiner D Nevertheless, it is quite possible that by chance Examiner A's candidates might have been re-examined by Examiner D, and Examiner B's by Examiner C. Thus we are justified in placing them together in the same table.

Table 4 3 is easy to read. Look, for example, at the row marked "I" in the history chart. In the first column, headed "Number", we find that 8 students were awarded a First Class or First Division on the first marking. Out of these, only 1 was awarded a First in the second marking, 5 were awarded Seconds, and 2 Thirds. Or we can read the same

chart downward Here it is the first row (in column "I") that tells us that 15 candidates were awarded "I by this examiner Of these only 1 was given the same mark by the other examiner, 10 were awarded a Second, and 4 a Third (We have not labelled the two sides "first' and "second' examiners, as they could easily have been interchanged)

To make the charts simpler to read, boxes have been put around the squares which represent agreement between the two examiners At the bottom you will also find the percentage of the total group who were awarded the same Division or Class by both examiners This percentage ranges from 59 9% for history to 77 1% for geometry So even in mathematics, nearly one student out of every four would be awarded a different Class or Division if he were examined a second time

A critic who saw these charts suggested that we got only two-thirds agreement on Class or Division because it is easy for candidates on the borderline to go up and down one point A glance at the table of differences (Table 4 1) will show that it was not just changes of one point that produced these results Furthermore, anyone familiar with objective test reliability scatter charts, will know that agreements above 90% can easily be achieved with reliable examinations

Table 4 3 shows us that, taking all thousand candidates, *even in geometry an average of one quarter of the candidates would NOT be awarded the same Division by a second examiner* For other subjects the average number changed is more than one-third

Taking all four papers (4,000 candidates in all), Table 4 3 shows that, on the average, candidates are likely to receive the same Division from two examiners only two-thirds of the time One and one third per cent of them (13 out of every 1000) would have two marks *two Divisions* apart, if their answer books were examined a second time Note that even in mathematics, there were eleven candidates who were shifted from a First to a Third or a Third to a First, a Second to a Fail or a Fail to a Second Fortunately, none of the candidates was shifted three divisions Thus we may state that if a candidate is awarded a First by one examiner, we can at least be sure he is not likely to be Failed by another examiner But a Second may be a potential Failure—and vice versa—*in at least one candidate out of every one hundred*

Table 4 3 shows that even the meaning of "passing" and 'failing" is uncertain

62% of the candidates were passed in both markings
22% of the candidates were failed in both markings
16% of the candidates were passed once and failed once (or vice versa) in the two markings

Twenty-seven per cent of all those failed in either marking were passed in the other marking (This is based on the total failed, and not just

those who crossed the borderline)

Students of educational statistics will be interested in this The over all coefficient of correlation* for the Mathematics scatter chart shown here is + 93 Thus we have a high examiner reliability—utterly dissipated by the use of a scale with a standard deviation twice that of any other examination The fault of course, is not in the use of the entire

TABLE 4 3

SCATTER CHARTS OF THE CLASSES OR DIVISIONS AWARDED BY PAIRS OF EXAMINERS (BOTH SAME AND DIFFERENT) TO ONE THOUSAND ANSWER BOOKS IN EACH OF FOUR SUBJECTS

History

	No	Dist	I	II	III	Fail
Number	1000		15	178	426	381
Distinction						
I	8		1	5	2	
II	120		10	62	40	8
III	618		4	109	334	171
Fail	254			2	50	202

Same Division = 59 9%

Hindi

	No	Dist	I	II	III	Fail
Number	1000	1	10	146	447	396
Distinction						
I	20	1	5	12	2	
II	158		4	81	68	5
III	531		1	50	320	160
Fail	291			3	57	231

Same Division = 63 7%

* Correlation coefficients in pooled data may be either higher or lower than those in the original separate batches (Guilford 1956 pp 322 325) The median r s reported in Table 10 1 are more accurate indicators and therefore coefficients for Table 4 3 have not been calculated The mathematics correlation is for the purpose of illustration only

TABLE 4 3—(*contd*)

SCATTER CHARTS OF THE CLASSES OR DIVISIONS AWARDED BY PAIRS OF EXAMINERS (BOTH SAME AND DIFFERENT) TO ONE THOUSAND ANSWER BOOKS IN EACH OF FOUR SUBJECTS

Biology

	No	Dist	I	II	III	Fa l
Number	1000		50	351	418	181
Distinction						
I	73		24	43	6	
II	371		26	223	115	7
III	438			84	274	80
Fail	118			1	23	94

Same Divis on = 61 5°

Geometry Mathematics

	No	Dist	I	II	III	Fa l
Number	1000	134	139	152	164	411
Distinction	148	119	29			
I	143	15	94	30	4	
II	140		15	94	28	3
III	176		1	25	103	47
Fail	393			3	29	361

Same Division = 77 1%

Average Same Division = 65 5%
Within one Division = 33.2%
Two Divisions d fference = 1 3%

scale, which is theoretically a good thing The fault is in the imposition of a system of Classes or Divisions from the scale used in most arts and science subjects on a totally different scale used for mathematics This is approximately as logical as expecting water to boil at 100 degrees F just because it boils at 100 degrees C There are two possible solutions to this (1) Pass and division marks may be set differently for mathematics than for other subjects (In many U S A universities where 60% is the

pass mark in most subjects 70% or 75% is required for a pass in mathematics) (2) Alternatively, statistical methods can be used to equate or calibrate mathematics marks to the more usual scale (See chapter on Scaling)

The remaining charts may be examined by the interested reader *On the average one-third of the candidates were placed in different Classes or Divisions in the two markings* It is obvious that we cannot place a very high faith even in Class or Division

5 What are the chances of a change in Division if an answer book is marked again?

There is another way of looking at the data in Table 4 3 What the student (and his teacher, who is not his examiner) wants to know, of course, is this "Suppose my answer book were marked again what are my chances of having my class or division raised or lowered?" Figure 4 1 answers this question

To get a reasonably stable answer, we used the pooled data for 1000 answer books in each subject from Table 4 3 But since there is really no logical reason why one examiner should be considered the "first' marker and another the 'second'—i e for, logically, they might as easily be reversed—we averaged the results of the two *

Figure 4 1 gives the results of this study For example, look at *History*

The first graph shows that there were 12 First Divisioners in history When these were marked again only 9% of them remained Firsts In other words, only 9% of these First Divisioners were awarded a First by *both* examiners Sixty five per cent were lowered to a Second Division by one of their two examiners, and 26% were awarded a Third

It is quite apparent that a "First" in History usually means only that one particular examiner is pleased—not that there is any inherent merit in the answer book that would be recognized by others also

The second graph shows that out of 149 Second Divisioners marked again, 5% went up to a First and 3% were Failed And even of the

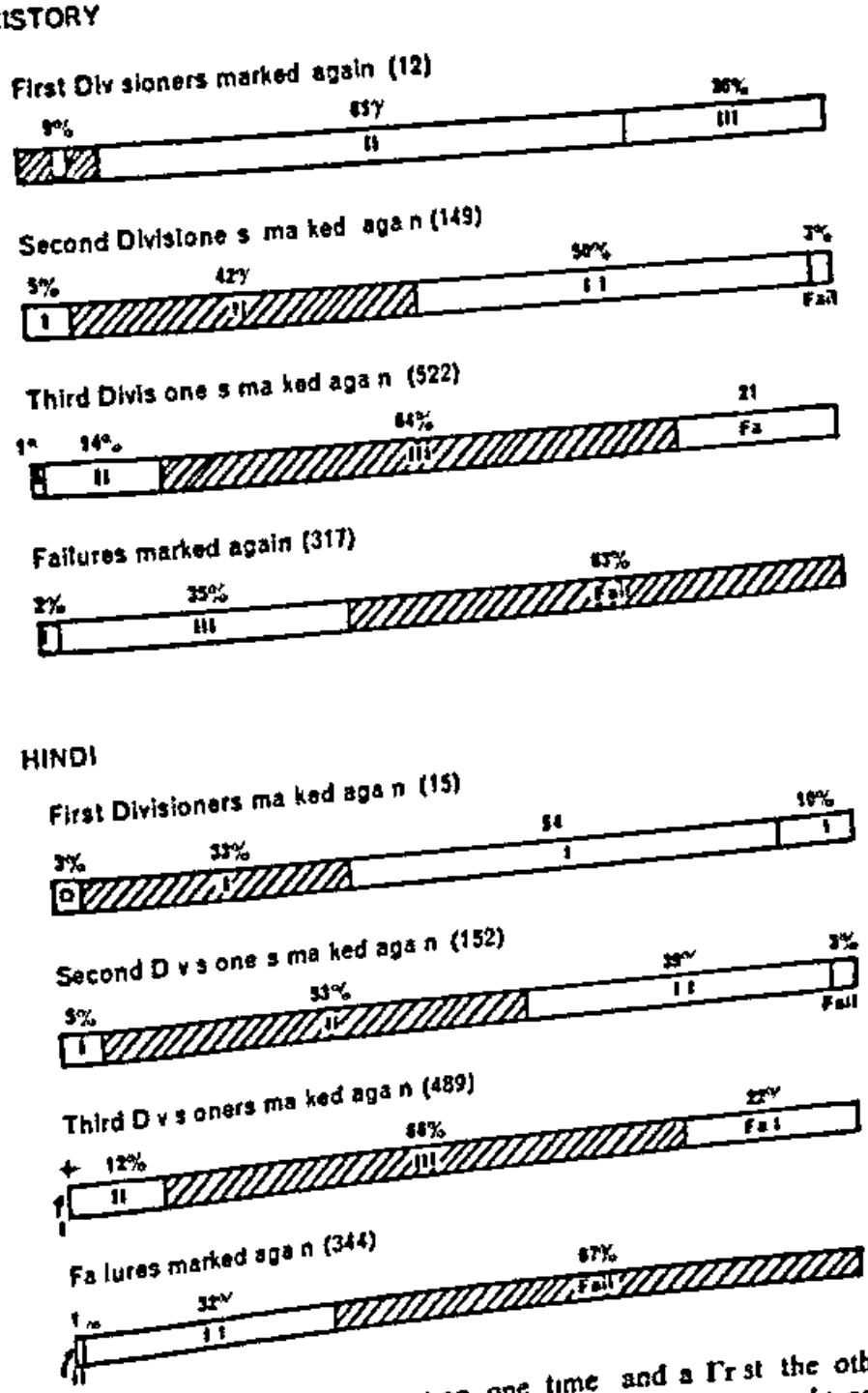
* Take for example the graph for geometry First Divisioners From Table 4.3 we find that one examiner awarded 143 Firsts while the other awarded 139 Firsts They are averaged as follows

Number	Dist.	I	II	III
143	15	94	30	4
139	29	94	15	1
282	44	188	45	5
100%	16%	66%	16%	2%

The average number of First Divisioners is 282/2=141

FIG 4 1

Can we trust Class or Division? The graphs show how many Cand dates would rece ve different D vis ons if answer books were marked aga n (Number)

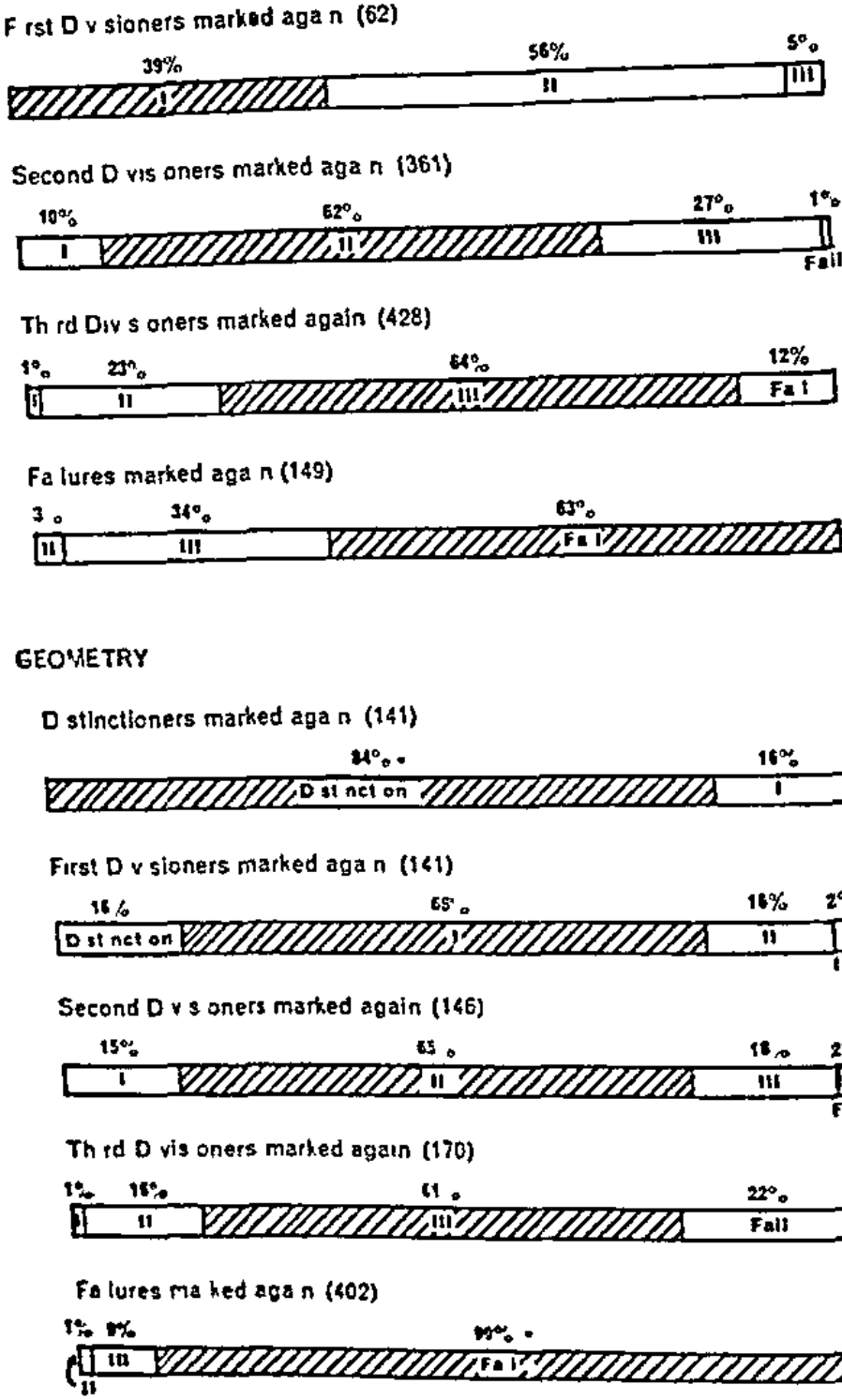


* One cand date was marked a D stinct on one time and a Fr st the other time This also accounts for the 3% chance (1 our of 30) that a First m ght go up to a D st nction

+ Three Third Division cand dates out of a thousand (0 3%) go up to a First Division This thus too small to graph—but substantial when we cons der the large number of Thirds usually awarded

FIG 4 1—(*contd*)

BIOLOGY



GEOMETRY

* Do not be m sled by the fact that the categories of D stinction and Fail *seem* to be more reliable than the others They are not There is nearly twice as wide a range of marks in the 'Distinction category and nearly three times as wide a range in the Fail category as in First Second, or Third Thus a Third Divisioner s mark needs to be changed by only 7 (out of 50) to put him in another Division—while a Fa lure s mark can be changed as much as 16 points and still remain a Fa l

Failures, at least 37% would have been passed by a second examiner, and two out of each hundred Failures awarded a 'Second

The results for *Hindi* show somewhat more agreement than those for history One-third of one per cent of the Thirds were awarded a First the second time in Hindi, as against 0 6% in history

Biology examiners show a substantially higher degree of agreement than history and Hindi examiners in the First and Second Divisions Still, the probability is *more than 60%* that any First Divisioner will *not* be awarded a First by a second examiner, or even by the same examiner going through his answer book a second time

	D	I	II	III	Fail
History		9%	42%	64%	63%
Hindi		33%	53%	66%	67%
Biology		39%	62%	64%	63%
Geometry	84%	66%	65%	61%	90%

Geometry examiners show a much higher degree of agreement for the First Division, but about the same as biology for the Second and Third Divisions The above summary table of 'agreement both times the anwser book is marked" is instructive

The figures for Distinction and Fail for Geometry are misleading The Distinction category has nearly twice as many "marks in it as I II or III, (which, incidentally, also differ from each other) while Fail has twice as many "marks" as I, and nearly three times as many as III (Would not a more rational scale allot equal numbers of marks to each category?) As we will see from Figure 4 4 and Table 4 5, Geometry uses *all* the marks in the Distinction and Fail categories, in a nearly flat curve —while other subjects use only some of them, in steeply declining curves This fact alone would explain the high degree of agreement found for geometry, for these categories Thus the bottom 100 candidates in geometry (see Table 4 5) could be raised 13 marks each and still be classified as Fail But if you raised the bottom 100 candidates in history by 13 marks each, only 5% of them would still be classified as Fail

It is evident from Figure 4 1 (as we also noted in a previous section) that *the only difference on which examiners always agree, is the difference between First Division and Failure* Thus we can feel sure that a First Divisioner is better than a Failure—but about Seconds and Thirds we can draw no clear conclusions

It must be added, however (as is discussed elsewhere), that the results on a single paper are less reliable than the results on a two-paper subject We can expect somewhat higher agreement on a subject of 100 marks, made up of two papers, than on either 50-mark paper However, the agreement will not be *very* much better The above conclusions still

stand—*a candidate's "Division" is as much a matter of the chance of who marks the answer book, as it is a matter of the inherent ability of the candidate.*

Let us suggest, however, that students look rather carefully at Figure 4.1 before they demand that their answer books be re-examined. Except for Failed students, in every other category (with the single exception of biology Third Divisioners), there is *more chance that the mark will be lowered than that it will be raised*, on re-examination. (Statisticians will recognize this as a good example of the familiar principle of "regression towards the mean".) Thus re-marking—if it is ever permitted—will help the low-mark students only, and the student who hopes to "improve his division" is not likely to benefit.

A closer look at the Failures is in order. Notice that 37% of the Failures in history and biology, and 33% of the Failures in Hindi were passed by a second examiner. (Only 10% of the geometry Fails were changed, but as we have seen this is partly an artifact of the extremely flat marking curve.) Educators may well ponder the value of an examination system where *one-third of all "Failures" in a subject would be passed on re-examination.*

6. *Six examiners compared on equivalent answer books*

There may be a suspicion that some of the results reported are due to the fact that different examiners received different batches of answer books. Perhaps the batch influences the examiner as much as the examiner influences the batch. Suppose, as in the ninety examiner project, all examiners had marked the same batch of answer books?

TABLE 4 4

LARGEST DIFFERENCE BETWEEN MARKS OF ONE HIGHLY EXPERIENCED EXAMINER (X) AND SIX OTHER EXPERIENCED EXAMINERS

History (50 marks)		Hindi (34 marks)		Biology (50 marks)		Geometry (50 marks)	
E	10	E	4	E	9	E	11
F	10	F	7	F	8	F	14
G	7	G	6	G	6	G	11
H	16	H	7	H	9	H	13
I	8	I	4	I	7	I	14
J	11	J	8	J	8	J	9

Reproducing 100 answer books would have been impossibly expensive, and besides we would have lost some of the other advantages of this particular study. However, in one part of this study we drew (in

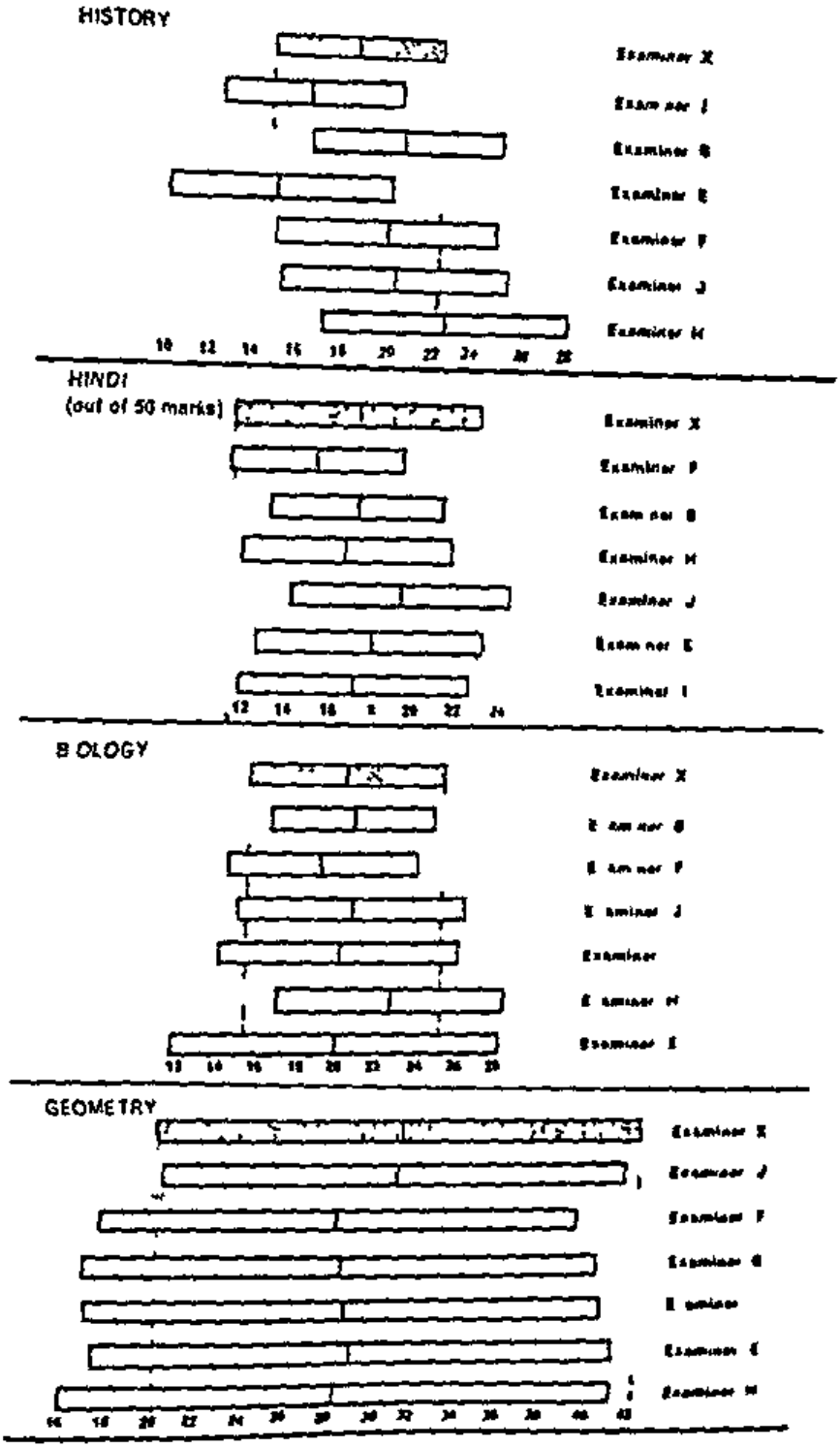


FIG 4 2

For each of the subjects we took 300 answer books which had been marked by one specially experienced exam ner and d vided them into six sets with equal (as nearly as poss ble) means and distributions. Each set of 50 was re-marked by a d fferent examiner The following are the results arranged in order of standard deviation Each bar shows the range of plus and minus one standard deviation (approximately the m ddle two-thirds of the marks) The line in the middle is the mean The shaded bar shows the results of the original examiner (Examiner X) the white bars show the s x re-examiners of the same answer books. (See also Table 10 3)

each subject) 300 answer books from one especially experienced examiner (called Examiner X) These answer books were divided into six equal batches The average marks in each batch were exactly the same Each batch of 50 answer books was sent to a different examiner (Examiners E to J, but a different six for each subject, of course) along with 50 of his own, making a total of 100

Figure 4 2 shows the results Start with the results of the seven history Examiners The top graph (shaded) shows the marks of the original specially experienced examiner, Examiner X, for all 300 answer books The line in the middle shows his mean or average marks The length of the graph line gives an indication of his distribution of marks (This is one standard deviation on each side of the mean)

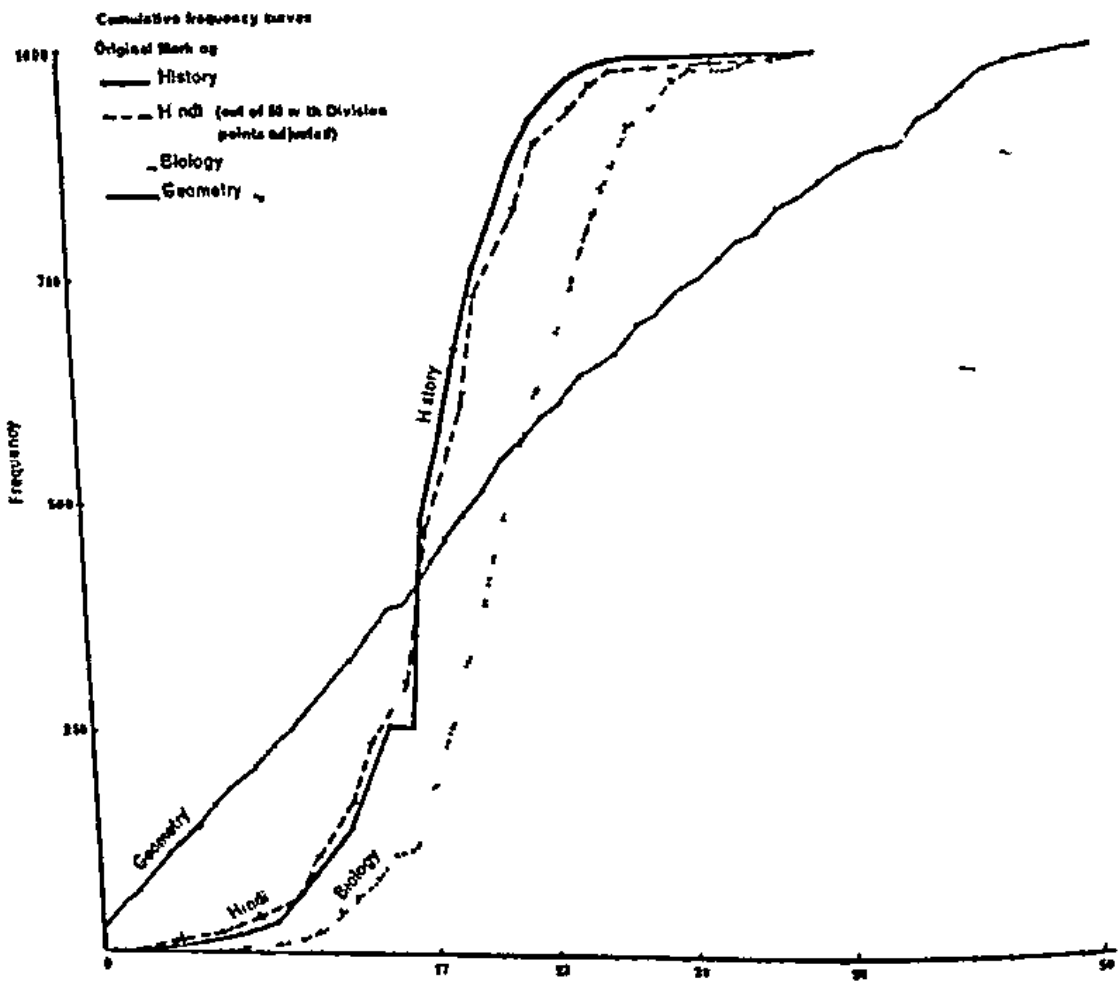


FIG 4 3

Cumulative frequency curves of marks awarded to four thousand Higher Secondary Board answer books (one thousand for each paper) by eleven examiners, during the Board Examinations of 1964 (Note Hindi marks have been multiplied by 50/34 and the marks from 7 to 34 have been raised one point each, so that the "Divisions" will come at the same place on the graph for all subjects)

The graph lines of the other experienced examiners (E to J) are arranged in order of length, the shortest being first Thus history Examiner I had the narrowest distribution of marks of all six But notice that his was wider than the 'standard", Examiner X In fact, all the experienced history examiners had a wider distribution than that of specially experienced Examiner X

The average marks awarded by Examiner I were however, much lower than those of Examiner X But look at the next graph line The average marks awarded by Examiner G were much higher than either those of X or I

The other graphs may be read in the same way The differences in mean (average) are much larger for history than for Hindi, biology or geometry But the differences in distribution are largest for biology

Notice, also, how very much wider the distributions of marks are for mathematics than for any other subject You would find the same sort of difference if the weight of one group of students were measured in pounds and that of another in kilograms But would you add these two together, to determine the average weight of the school? Adding Hindi and geometry marks is just as illogical, when the scales used are so different

Table 4 4 shows the largest differences between examiners Again, Examiner X is taken as the standard Notice that experienced examiner H differed by as many as 16 marks from Examiner X on the merit of one of the candidates For Hindi, the largest difference was 8 marks out of 34, which is equivalent to 12 marks out of 50 The experienced biology examiners were most consistent, but even here two of them—examiners E and H—differed by as many as 9 marks with the 'standard But the widest differences were in geometry, where two experienced examiners —examiners F and I—differed by 14 marks out of 50, on at least one candidate each, from especially experienced Examiner X

Thus, *even when experienced examiners receive equivalent batches of answer books, they differ* in average marks, range of marks awarded, and the merit of individual candidates

7 *Distribution of marks*

Table 4 5 reports the total frequency distributions of the marks in each of the four subjects The lines across indicate the dividing lines between the various Divisions O and R refer to Original marking and Re marking Hindi marks which were originally out of 34, have been translated into "marks out of 50' so that the distributions can be compared

Figure 4 3 gives the cumulative frequency curves for the original

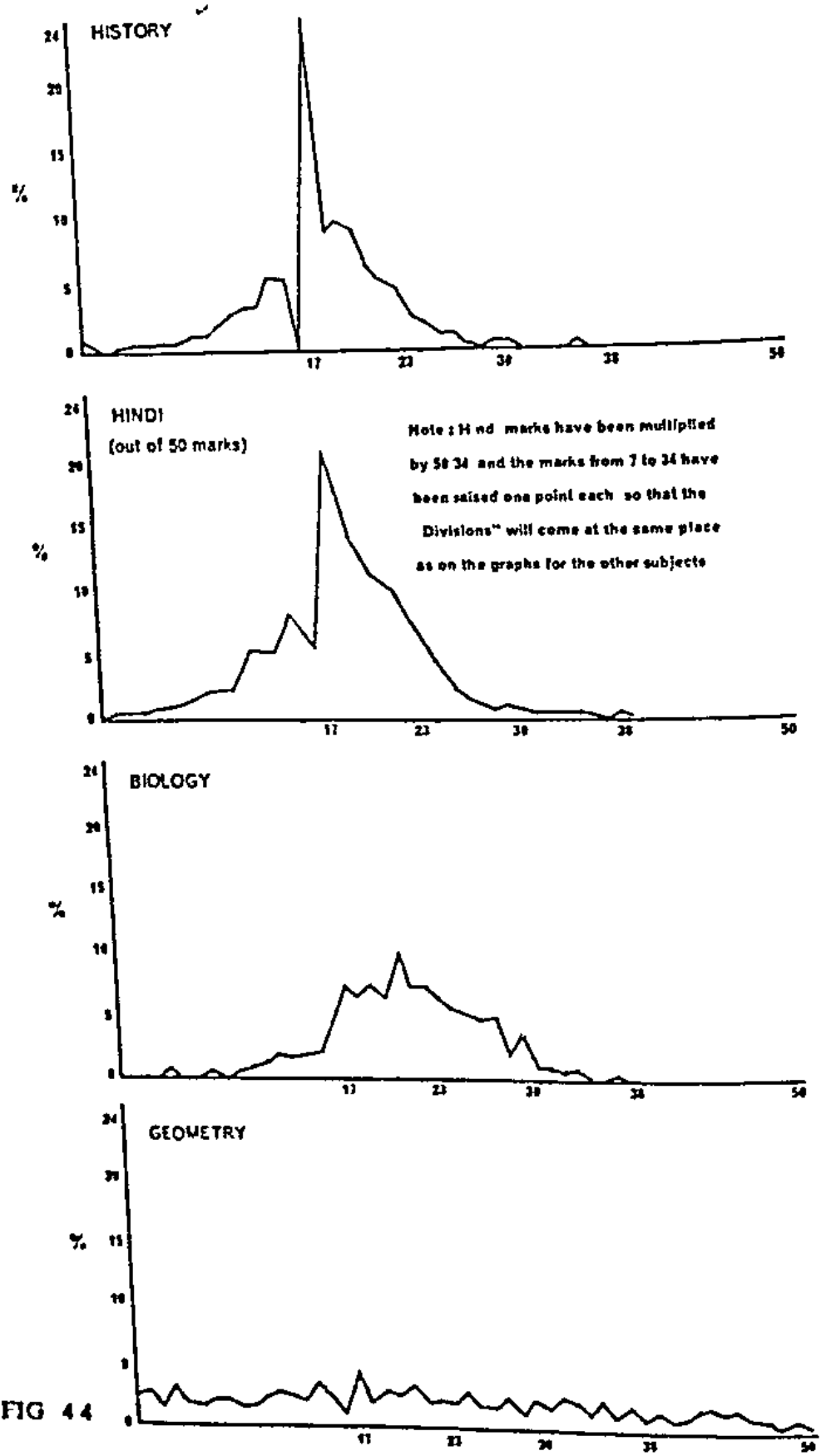


FIG 4.4 Distribution of marks awarded to four thousand Higher Secondary Board answer books (one thousand for each paper) by eleven examiners, during the Board Examinations of 1964

(first) marking, for the four subjects And for those not used to interpreting cumulative frequency curves, Figure 4 4 graphs the four distributions of marks In both figures, the lowest mark of each Class or Division is shown (17, 23, 30, 38)

It is interesting to see how very different the marking system seems to be, for the various subjects How else can we explain the very different distributions of marks? In this study, the lack of a completely random sampling may have contributed to this But other studies have shown the same differences Some may say, of course, that the different distributions represent differences in the abilities of the students, not different marking systems Although it is probably impossible to prove definitely that this is not so, it is certainly highly unlikely Harper (1963b), for example, showed that students with the same ability received different ranges of marks in biology and mathematics Mahalanobis (1934), and others, have shown widely different distributions in different subjects for the same sample of students The Education Commission (1966, page 291) noted very pointedly that "an 80 per cent mark, say, in mathematics, does not convey the same meaning as, say, 80 per cent mark in history or English "

This would not matter if we treated each subject's marks separately, and did not add them However, when marks are added, the distributions affect the actual weight that each mark carries in the total (See Mahalanobis, 1934, or any standard book dealing with the subject Taylor, 1963, also emphasized this point) The unstated assumption in aggregate marks seems to be that each subject carries equal weight (Or, rather, that the weight of the subject in the aggregate total is determined by the number of marks allotted to it) This is simply not true when unscaled marks are added The result is that some subjects—for example mathematics—may actually be twice or three times as influential in determining the candidate's final rank as are other subjects which are supposed to be equally important

It is also noteworthy that these curves are obviously not "normal" Both theoretical and factual evidence indicates that abilities are normally distributed in any average group The progressive selection of students that has taken place (the weaker ones dropping out before reaching high school exams) can explain a slight decrease in lower scores (negative skewness) But there is no such way to explain the flattening (platykurtic nature) of the curves (Yes, they are platykurtic, if you average out the artificial abnormalities around the passing point) This is, of course, especially striking for geometry Some teachers claim that the methods they use to determine marks are "natural" and that any statistical scaling procedure would be imposing an "artificial" scale on this natural situation The above facts, however, prove exactly the opposite The methods

TABLE 45

FREQUENCY DISTRIBUTIONS OF MARKS AWARDED BY ALL EXAMINERS IN FIRST AND SECOND MARKINGS OF FOUR THOUSAND HIGHER SECONDARY BOARD ANSWER BOOKS (11 ORIGINAL AND 10 RE EXAMINERS FOR EACH SUBJECT)

Subject	History		Hindi		Biology		Geometry	
Original or Re-marking	O	R	O	R	O	R	O	R
Marks out of fifty: 50							5	4
49							9	4
48							5	6
47							6	9
46							10	9
45							17	11
44							16	14
43							19	16
42							18	12
41							9	12
40							7	17
39							16	10
38			(26)	1			11	10
37			(25)	2	1	1	18	20
36	1	1				4	13	10
35			(24)				23	16
34		1	(23)	1	3	7	1	13
33		2				8	4	18
32		1	(22)	3	1	10	7	18
31	3	2	(21)	3	1	12	15	24
30	4	8				35	17	15
29		6	(20)	11	5	19	34	19
28	4	19	(19)	7	9	47	20	13
27	13	21				47	46	26
26	11	29	(18)	17	14	71	58	15
25	21	30	(17)	24	30	55	61	16
24	25	41	(16)	42	38	61	60	28
23	45	32				71	72	18
								22
								19

(Marks out of 34)

TABLE 4 5—(*Contd*)

22	53	41	(15)	68	55	73	67	21	26		
21	60	59	(14)	92	87	96	85	32	26		
20	87	61				62	77	27	25		
19	93	59	(13)	112	105	72	63	30	29		
18	88	81	(12)	133	121	65	61	22	23		
17	237	125				70	66	46	35		
			(11)	194	134						
16		20				8	32	8	14		
15	50	69	(10)	52	77	23	33	27	27		
14	52	68				22	26	32	24		
13	32	37	(9)	74	87	19	23	20	21		
12	31	41	(8)	48	84	21	10	25	37		
11	29	39				11	19	28	21		
10	20	25	(7)	52	46	7	10	24	17		
9	12	36	(6)	22	47	3	13	21	23		
8	10	11					9	20	23		
7	6	15	(5)	20	29	3	4	25	26		
6	3	5	(4)	11	13		2	23	18		
5	4	4						18	18		
4	1	6	(3)	8	9	1		22	27		
3	2	1	(2)	3	2			31	24		
2	1	2						13	22		
1		1	(1)	1	2			29	33		
0	2	1						27	36		
	1000	1000		1000	1000	1000	1000	1000	1000		
Number	17 77	17 59		17 17	16 43	21 85	21 00	21.25	20 62		
Mean				(11 68)	(11 17)						
				5 009	5 191	5 121	5 372	13 420	13 327		
σ	4 459	5 617		(3 406)	(3 530)						
					1%			14 8%	13 4%		
Distinction	—	1 5%		2 0%	1 0%	7 3%	4 9%	14 3%	13 9%		
I	8%	17 8%		15 8%	14 6%	37 1%	35 1%	13 8%	15.2%		
II	11 9%	42 6%		53 1%	44 7%	43 8%	41 9%	17 8%	16.4%		
III	61 8%	38 1%		29 1%	39 6%	11 8%	18 1%	39.3%	41 1%		
Fail	25 5%										

Note —"Distinction" row, for Mathematics especially

now being used to determine marks (particularly in mathematics) are actually quite artificial and arbitrary Furthermore, they take no account of all we now know about the distribution of human abilities and achievements This situation can only be corrected by the application of scientific scaling techniques to examination marks

Note what Dr Taylor calls "the J-effect". This is the piling up of "just barely passing" marks You can see in the graphs what the late Dr Habibul Rahman, in his usual delightful way, referred to as "the Qutub Minar' effect In history, for example, 237 candidates are 'awarded' (and this is the only appropriate word, since this obviously is in no rational sense a measurement) a mark of 17 (passing), *none* are awarded 16 (which would be just barely failing) This is a sort of educational 'brinkmanship" for which there is very little scientific or rational justification At best, it is an obvious confession on the part of the examiner that he knows his results are fallible—and he does not want anyone to fail 'by chance" Yet, if this is the motive, (and it is certainly a laudable motive) there are *far better* ways to ensure this result (See, for example, Taylor's 1964 discussion of "Grace Marks", which is summarized in Appendix B of the present volume) Actually, by trying to adjust just the borderline marks, the examiner takes care of only a small fraction of the total range of "failure by chance" marks, anyway Furthermore, a simple knowledge of correlation between examinations (generally low) and what that implies would indicate that these tactics are not likely to be effective in many cases, anyway How much better to be honest, and eliminate this "J-effect' in marking

The fact that examiners would probably *like* to be more honest is indicated in Table 4 5 by the second column, R, of each pair Notice how *greatly reduced* the J effect is, when the examiner marks the exam books the *second* time Knowing that the marks "will not really count", he apparently feels freer to be completely honest in marking The number for whom he feels justified in adding an extra mark "to keep them from failing' is much smaller under this circumstance Yet even in these circumstances, the unadmitted assumption that 17 is "pass" remains To do a completely unbiased marking, apparently, one must cease to think in terms of "pass" or "fail'

The answer to this problem is simple (1) Introduce the scaling* of examination marks and then (2) completely outlaw the J-effect in marking Inform the examiner clearly that the examiner s "17" is no longer "passing', because scaling may change it If an examiner's marks are piled up at the borderline, these borderline answer books should be returned for re marking It would be simple to send only these, asking

* See the chapter on Scaling in this book.

him to lower the marks of half of them by 1 mark (Perhaps, after doing this once, the examiner may be cured of the habit) This procedure will be quite necessary if examination marks are scaled—because scaling will almost invariably push this J into some quite irrelevant position Would it not look silly to find 237 receiving 19 marks, and no one receiving 18?

Notice that the J-effect is largest in the subject which is marked most completely on the basis of judgement—history Of course, it is the examiner's distrust of his own judgement that produces the J in such subjects But it is not absent from the other more "objective" subjects, by any means

At the very bottom of each column of Table 4 5 is given the percentage receiving each Division Notice that the percentage failed increases sharply when the papers are re-examined, in each case This is due to two factors (1) the drastic reduction of the J-effect, and (2) the increase in the standard deviation (distribution) of marks

The differences in percentage in each Division, among the various subjects, confirms what we have said above about the very different marking systems used (See Harper, 1963b) Is it fair that 13% of Mathematics students should be awarded Distinctions, while only 1 in 1,000 is awarded this in Hindi—and none in history or biology? Conversely, does it seem scientifically correct that two or three times as many should be failed in mathematics as in biology? We have heard some mathematicians say, "it is the nature of the subject' It is not—it is the nature of the marking system, the method and instructions given to the examiners, not of the subject If teachers of English or biology also gave full marks for correct answers, their marks would be much more like those of mathematics There is ample evidence for this, as, for instance, the fact noted earlier, that students of the *same* ability receive *different* marks in different subjects (See Harper, 1963b) In any rational and scientific marking system, students of the same ability would (on the average) receive similar marks in all subjects

Hindi presents a rather odd and interesting case, also Here, somehow, the pass marks are 32% in this paper, rather than 33% ! Eleven marks (passing) is 32% of 34 marks Yet a student achieving 11 marks on each of the three papers (for which the maximums are 34, 34, and 32) will receive a total of 33 out of 100, and will pass Of course, "passing", "failing", "Divisions", etc are only calculated on the total, not on individual papers Yet the examiner obviously has some sort of a "pass mark" in mind when marking We instructed the computer (an IBM "1920"), in tabulating these results, that it was to consider 33% as passing The computer, being a completely logical machine, "passed" only those with 12 marks (35%) and "failed" those with 11 marks (32%) But when we found a reversed J-effect with 133 passing and 194 failing, we realized

that human examiners must be following a different system. The larger group of 194 must be passing, not failing. This is why the lines across, between Divisions, are one mark lower in the Hindi columns than in the others. However, in order to make the graphs (Figures 4.3 and 4.4) visually comparable for all subjects, we raised the Hindi marks between 7 and 34 (inclusive) by one point each.

8. *Reliability*

What we have discussed so far is just the difference between two examiners marking the *same* essay-type examination. But suppose the candidate also takes a *different* examination, even one which is supposed to measure exactly the same subject-matter as the first examination? In other words, suppose he takes two alternative examinations, each marked by a different examiner? How different will his marks be?

The over-all summary of this study is contained in Figures 4.5 and 4.6. Figure 4.5, the coefficient of reliability, will be readily understood by the technically-trained. But Figure 4.6, the Standard Error of Measurement, is in many ways more meaningful, and is certainly easier to explain to the layman.

There are errors in *all* measurements. (Try asking ten men to measure the length of the same table. You will get ten different answers.) By certain mathematical formulas we can describe this "margin of error" for an examination. This figure is known as the standard error of measurement (SEM).

Suppose that a student was examined 100 times, on a hundred different but equivalent papers and marked by a different examiner each time. (We can't do this actually, of course—he would go on strike!) We then might think of the *average* of these 100 different marks as the "true mark" of the student. But of course, some of his marks would be above this true mark, and some below it. Two standard errors of measurement (one above and one below the true mark) gives the range *within which two-thirds of the marks will lie.* Four standard errors of measurement (i.e. plus or minus two) give the range of about 95% of the marks. If we take a range of six standard errors of measurement, virtually all of the candidate's possible marks will lie within this range.

We can put this a little differently. If 95% of the marks lie within a range of plus or minus two standard errors of measurement, we can also say "There are five chances in 100 that the student's mark will be *more* than two standard errors of measurement above or below his true mark." From Figure 4.6 we find the average standard error of measurement of geometry marks is nearly 5.4 marks out of 50. Twice 5.4 is 10.8, or 11 approximately. Thus there is a 5% chance for any candidate that

his mark on a mathematics examination will be 11 or more points (out of 50) higher, or 11 or more points lower than his 'true mark This also means that we can expect this much deviation for 5 out of every 100 candidates Suppose, for example, we have 100 mathematics candidates of exactly equal merit, who would all receive 27 marks out of 50 on a perfectly reliable test Ninety five per cent of these will be between 16-38 marks But five per cent of these equal merit candidates will be either *above 38* or *below 16* marks out of 50 In other words, out of every 1,000 who should all be Second Divisioners in geometry, *there are 25 who receive Distinctions and 25 who Fail*

From Figure 4 6 we see that the situation for history, Hindi, and biology is not quite so bad Here the averages range from 3 2 to 3 4 marks Thus, for 100 biology candidates who should get 27 marks out of 50, 95% will receive marks within about 7 marks above or below their "true marks Two and a half per cent will get marks of 34 or more, and 2 1/2 will be marked 20 or less Thus they will range only from First to Third Division

To make this clear, the data are presented in a graphic fashion in Figure 4 7 Here two ranges are plotted plus and minus one standard error of measurement (which includes 68% of all errors in marks) and plus and minus 2 58 SEM (which includes 99% of all errors in marks) The average SEM of each examination has been graphed At the bottom of the graph is the range of a rather poor objective type examination (Assumed reliability 85, with standard deviation the average of history and biology) As can be seen, objective examinations are not without error—but their error is considerably less than even the best of the four essay-type examinations Note that the error factor in mathematics is about half the total range of the 50 marks used

As the reader will discover later (and the technical expert has already noted from Figure 4 5) mathematics is actually *more* reliable than the other examinations Why, then, are the marks themselves apparently *less* reliable than those of the other subjects?

The reason is because Mathematics examiners *use a different scale from the scale used by most other examiners* (See Harper, 1963 b Mahalanobis, 1934) Comparing mathematics and biology marks is like comparing centimetres with inches The marking system used in mathematics has a much smaller unit than other examinations Each mark in mathematics is apparently worthless than a mark in almost any other subject Therefore, a much larger number of them is awarded to describe a student of high ability (just as more centimetres than inches are "awarded" to a long table) But when we then treat mathematics marks as being equal to other kinds (as in Figure 4 4 or Table 4 5, or in awarding Classes or Divisions) the result is a much greater number of Distinctions

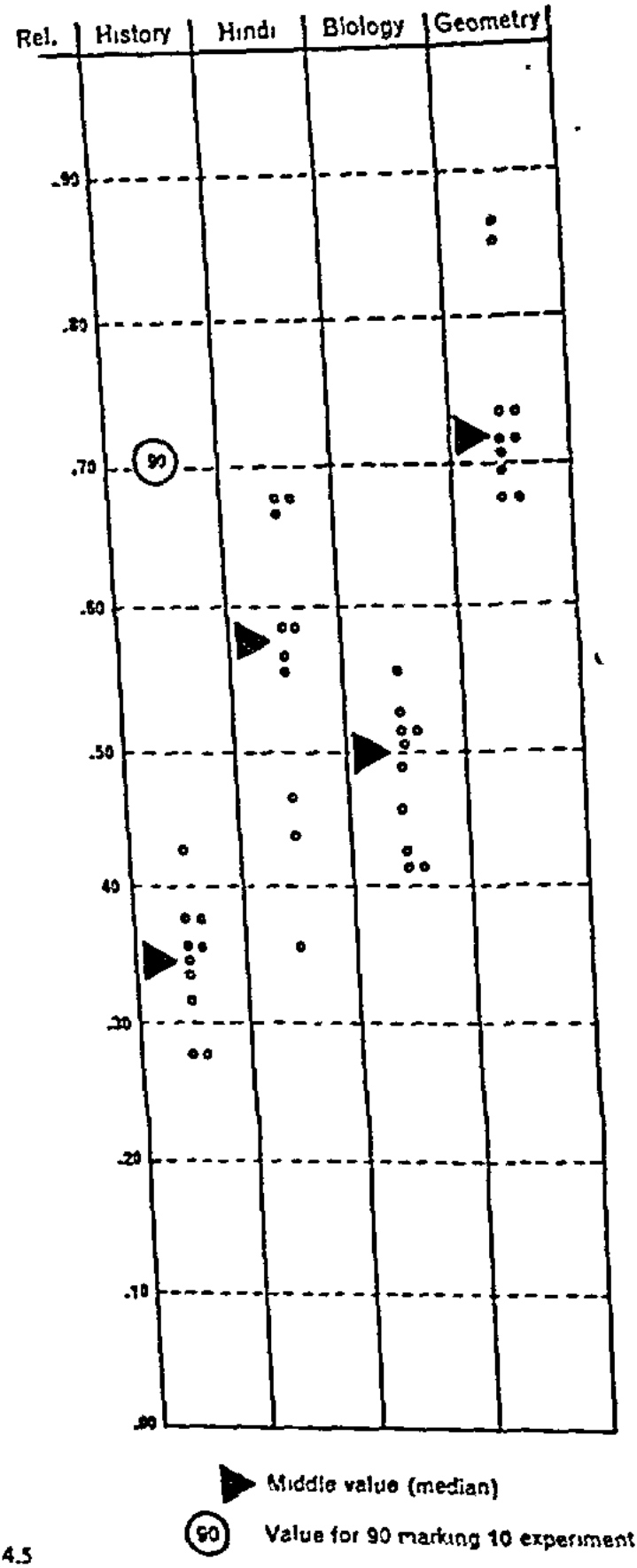


FIG 4.5

Coefficients of "*marks reliability*" for ten sets of fifty answer books marked by pairs of different examiners, in each of four subjects

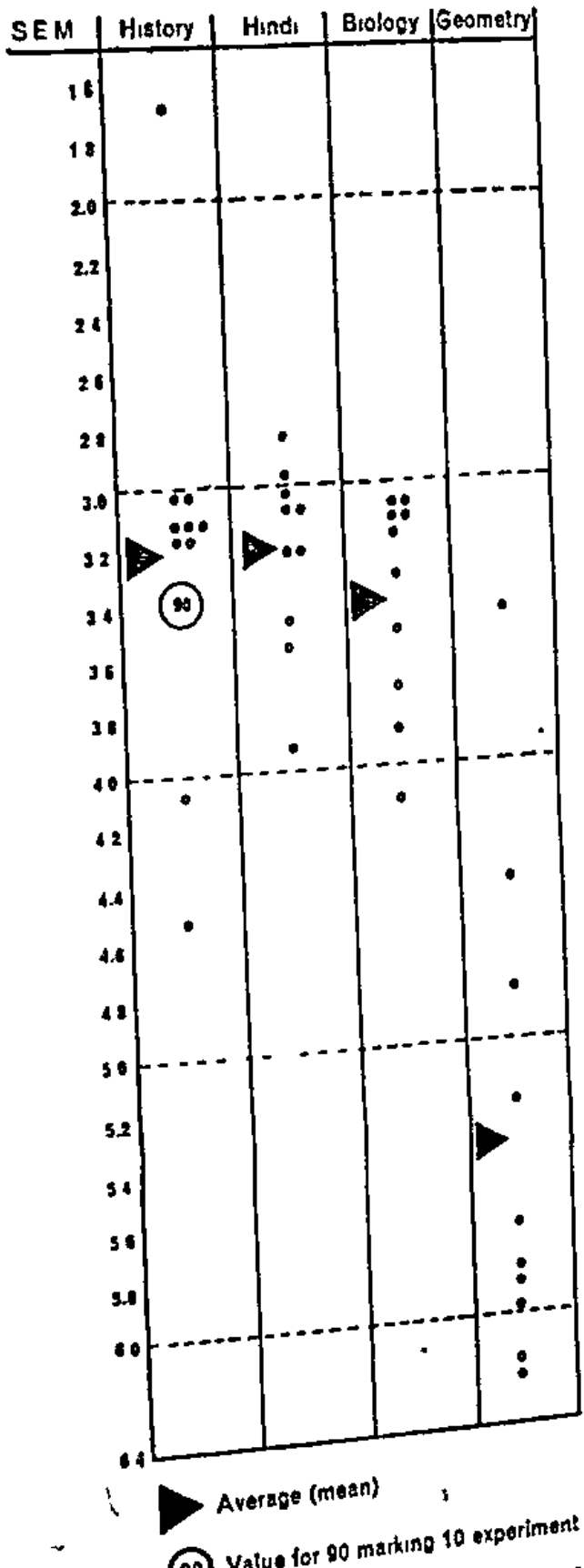


FIG 4 6

Standard Errors of Measurement for ten sets of fifty answer books marked by pairs of different examiners, in each of four subjects (Marks out of 50)

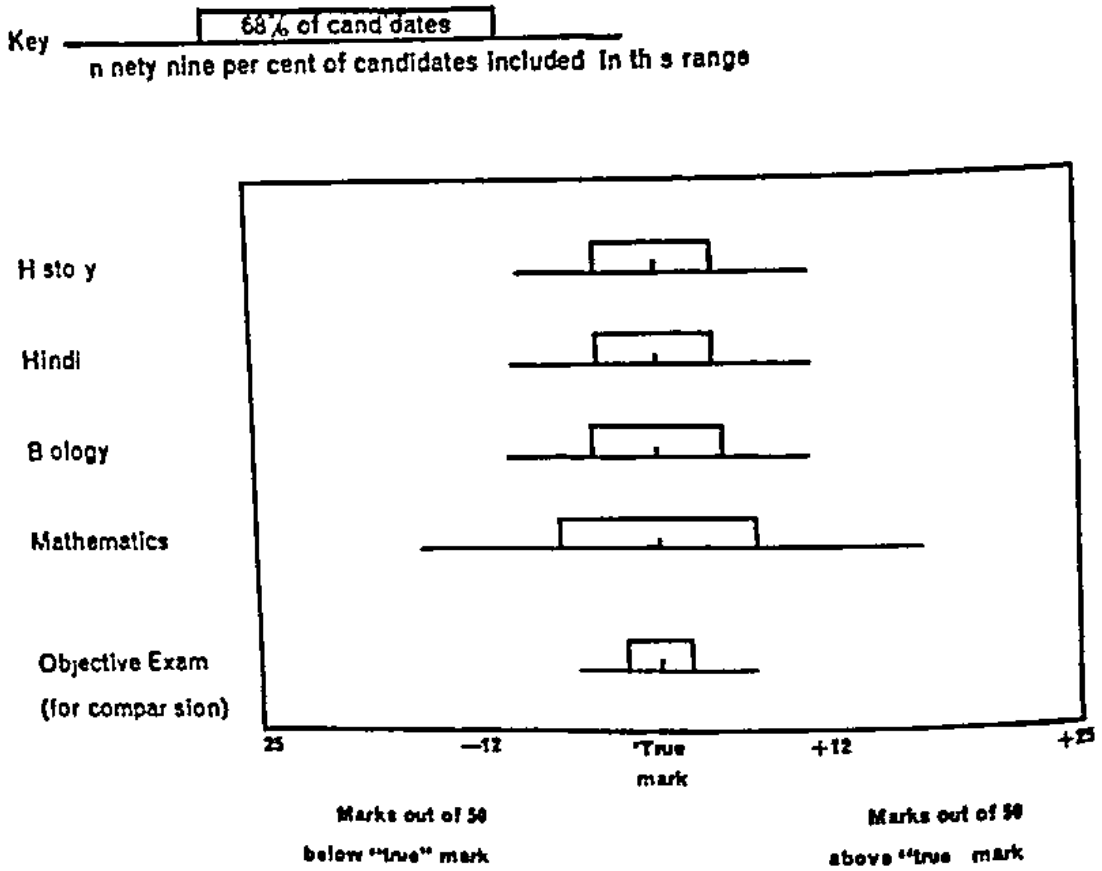


FIG 4 7

Relationship of error in examinations to the total range of marks out of 50, based on average Standard Error of Measurement of each examination

and of Failures than is really justified And the result, also, is that the SEM is unnecessarily enlarged and the reliability (in the layman's sense) of marks is lowered

If — as has been widely advocated, but seldom done in India — all marks are "scaled before they are published, then the advantages of better marking in mathematics will show up in a smaller standard error of measurement As a matter of fact, if this were done the range of error in Figure 4 7 would be only slightly larger for mathematics than for the objective examination By "scaling ' we mean that all marks are changed, statistically, into the *same* scale, rather than there being different scales for different types of subjects (Mahalanobis, 1934, Gulliksen 1960)

Now a further word about the reliabilities in Figure 4 5 The reliabilities graphed here are ' marks' reliabilities You will find later (Chapter VII) a more complete discussion of this concept Briefly, we have three types of reliability reader reliability (the correlation between two examiners marking the same answer book), content reliability

(analogous, in an essay test, to "reliability" in an objective test), and marks reliability If a student takes Form A which is marked by Examiner X, and then *strictly parallel* Form B which is marked by Examiner Y, the correlation between the two sets of marks will be the "marks reliability" of the test As can easily be seen, this is the *only* essay examination reliability figure which can be compared directly with objective examination reliability Kelley's famous standard* suggested a reliability of 94 for individual measurement If we are very lenient and reduce this to 85, we still find only two out of our 40 reliabilities reaching this standard

The conclusion is painfully obvious *The reliability of traditional examinations — even in mathematics — is not adequate for the assessment of differences in the abilities of individuals* This is not a new discovery Since it has been known for a very long time, one wonders why we *continue* to bless or curse students *for life* on the basis of *results which are often only a little better than chance* **

However, we are not yet finished with Figures 4 5 and 4 6 In each of these, a circled "90" indicates the statistic for the "ninety marking ten" experiment. It is particularly striking that the coefficient of marks reliability for that study is so very much higher than any other history reliability figure (71, while the next is only 43) This is undoubtedly due to the fact that in that study, we selected ten answer books in which the same five questions had been answered In the present

* "Making the assumption that for a test to be useful it must permit discrimination of a difference as small as 0 26 times the standard deviation of a grade group with chances five to one of being correct, Kelley arrives at the following as the minimum correlation for several purposes

(a) To evaluate level of group accomplishment .50
(b) To evaluate differences in level of group accomplishment in two or more performances .90
(c) To evaluate level of individual accomplishment .94
(d) To evaluate differences in level of individual accomplishment in two or more performances .98

It must be recognized, however, that these values are arbitrary, being derived from the above assumptions as to what it would be reasonable to expect a test to do in the way of discrimination between individuals and groups How low a reliability one is willing to accept in any given case depends upon the practical values which are involved in that particular case" (Thorndike, 1951, p 609)

** One of the communications from the sponsors of this project was addressed to me as "Director, Research on Extermination" There is perhaps more truth than humour in this slip of the typewriter There is little doubt that many careers have been "exterminated" before they had a chance to flower by the irrational vagaries of traditional essay-type examination marks We can only count ourselves fortunate that Einstein and Ramanujan had the courage to continue despite early "failure" in mathematics—and John Maynard Keynes was not put off by receiving his lowest marks in the Public Service examination, in Economics

study, however, different candidates were — as usual — answering different questions *All* experts contend that examinees should be allowed no choice of questions (The only exception is when questions are set in different alternative textbooks, in which case the choice should be "either-or") The comparison of the results of our two studies dramatically confirms the large loss of reliability when students are allowed to select any five out of ten questions

The SEM is, of course, related to the marking scale used The marking scale used by the 90 examiners seems to be somewhat wider than that of most of the 10 examiners (Possibly this is related to differences in marking ten vs 100 answer books) Thus the SEM for the "90 marking 10 study is slightly larger than in the present study. Also note one very small SEM, which was due to one examiner awarding only an absurdly narrow range of marks

Finally, remember what was pointed out earlier (Chapter III, Section 7) about the standard error of measurement What we have graphed is the SEM for marks on each examiner's own particular scale These would hold for another examiner using the *same* scale But examiners use different scales, with different means and different standard deviations Thus the actual errors — as we have seen from Table 4 1 — are much larger than these SEM's would lead us to believe

9 *Is the situation hopeless?*

From the above it is obvious that *the reliability of our traditional essay-type examinations is completely inadequate for the purpose for which we use the marks, i e to decide individual merit* Does this mean that all examination results are totally unreliable?

Not quite Fortunately, as we add together the marks of correlated papers, the reliability of the aggregate marks tends to build up This is because some of the errors are averaged out The candidate who happens to have had a very stiff examiner on one paper may, by chance, have a lenient examiner on another paper Some, of course, will have a stiff examiner — or a lenient examiner — on both papers So the errors do not completely average out, by any means — and for some candidates, the errors may not average out at all However, at least the total number of candidates who are seriously mis-marked is reduced, when we take the total aggregate marks ignoring marks on individual papers

Unfortunately — as we have mentioned before — there is no "averaging out" in the case of a failure In this case, the chain is only as strong as its weakest link *At the crucial level — the decision as to whether a candidate fails or passes — the aggregate results are just as unreliable as those of a single subject* As Taylor (1964) has shown, our present un-

scientific system of "grace marks" does *not* adequately correct for these errors (See Appendix B)

10 Are examiners influenced by the number of pages written?

Many students are convinced that examiners just count the number of pages and award marks accordingly We found definite evidence that this is *not* true It is obvious that the student who knows a great deal will write more pages than the student who knows almost nothing so *some* relationship between examination marks and number of pages written is to be expected And this is all we found — just a very minimum relationship — in all but two of our 44 examiners Even for the two the maximum correlation was 669 This means that the influence of number of pages on marks was only about 45% of the maximum possible influence (In technical terms, this was a 45% overlap in the variances) The average correlation was 37, which means only 14% of the maximum possible influence So if the examiners in this study were at all representative, we can say that *less than one out of twenty examiners is influenced significantly by the number of pages written Students should be informed of this* It would be a boon to examiners if students would stop filling their answer books with so much "padding" and trash

11 The need for scaling

The reader has already found that a recurrent theme in this book is "the need for scaling"

The need for scaling examination marks has been pointed out again and again in India (Mahalanobis, 1934 Harper, 1963 a & b, Taylor, 1963, 1964, Education Commission, 1966) And the data in these studies, also, repeatedly point to this need

What is "scaling?" Perhaps the simplest and most familiar example is the difference between the Fahrenheit and Centrigrade scales of temperature Suppose one thermometer in your room reads 104 degrees (Fahrenheit) and the other reads 40 degrees (Centrigrade) Since they both read in "degrees" you average them and find the average temperature is 72 degrees Are you justified in concluding that 72 degrees is a very comfortable temperature, so all your feeling of extreme heat is just "psychological"?

Averaging the marks in different examinations, with different scales, is just as unscientific (Harper, 1963 b, has discussed some of the "scales" in use in India in more detail)

The proper procedure for averaging Fahrenheit and Centrigrade

"degrees" is first to convert one scale to the other Here is the formula

$$\frac{9}{5}C + 32 = F$$

Thus, in the problem above, (9/5) (40) + 32 = 104 —so the two thermometers are actually reading the same temperature Similar formulas in the form ax + b can be set up to translate the scale of any examination (or any examiner) to the scale of another examination (or examiner) More commonly, we translate the scales of all examinations and examiners to some arbitrary scale which we have accepted as the standard This scale may be based on the average of the scales of all examiners and subjects, or it may be based on some other type of arbitrary decision

Any student of educational measurement knows one or two scaling methods Others have already been published in India (Mahalanobis, 1934, Harper, 1963 a) A fuller discussion of scaling, its rationale, and various methods especially suited to public examinations are presented in a later chapter entitled "Scaling". It is hoped that, even though it contains some technical details, every reader concerned with the final results of examinations will read it (It is not necessary to read the technical chapters in between to understand the chapter on Scaling)

PART TWO

Detailed Reports

detail treatment of things item by item (pl) the items or parts of a composite whole

—*Pocket Oxford Dictionary*

5

*Review of the Literature**

THE PROCESS of education includes three major divisions formulations of objectives designing education for the achievement of the objectives and assessing the outcomes of education The function of examinations is to provide such information as is needed for a precise assessment of the outcomes of education As such examinations are as old as or rather older than any system of education written Examinations are of many forms They may be performance oral or Written examinations may be classified as objective-type, very shortanswer type short answer type and traditional essay-type

The genesis of traditional essay type tests may be traced to the Chinese system of administration in vogue around 1115 B C The Chinese used the essay type tests for the selection of candidates for Government offices Each candidate was shut up in his narrow cell for days and nights and was required to produce compositions in prose and verse on the themes assigned to him (A detailed description of the Chinese system of testing is given by DuBois 1964) It was probably because of lengthy testing time and vast coverage of test material that as Ebel et al (1958 p 1502) put it ' the competitive examination system of ancient China helped it serve its primary purpose of providing men of ability for the service of the state ' The Chinese system of testing travelled though not in that strenuous form from China to the West By the 19th century, the

* This chapter is by the junior author and is adapted from his unpublished manuscript *A Critical Study of Essay-Type Examinations* a thesis presented for the degree of Doctor of Philosophy Gauhati University 1970

essay-type tests in the traditional form were well-established in western countries for the award of degrees and diplomas. (A concise narration of early history of examinations is given by Ebel et al. (1958.)

The use of the essay-type tests in India came as an inevitable consequence of the introduction of the western system of education. Indian universities were modelled on the lines of the British universities; therefore, they also adopted the examination system in practice in the U.K. In India, the essay-type tests were used on a large scale for the first time in 1857 by the universities at Calcutta, Madras and Bombay for the selection of students, and later for the awarding of university degree. The wide use of the essay-type tests at the university stage gradually led to the adoption of such tests at the lower stages of education also.

Over the years examinations have attained too great an importance in the West as well as in India. People have been branded for life on the basis of their examination results. Not only a man's academic but also his social status has often been judged from his degrees and diplomas. This has led many to seriously question the accuracy of examinations. One of the earliest studies exploring the efficiency of the essay examinations was made as early as 1888 by Edgeworth (1888) in the U.K. Other studies followed suit, of which those done by Starch and Elliot (1913) in the U.S.A., and Hartog and Rhodes (1935) in the U.K are very important. Thus in the early half of the 20th century, systematic research on examinations was well started.

Advances in various branches of sciences helped examination research a great deal Examination experts no longer remained satisfied with guesses. The scientific research approach came to be adopted for the study of almost every problem in examinations. In the beginning, attempts were made to minimise subjectivity in marking. Many kinds of objective-type tests were worked out. Various scoring methods were evolved to minimise chance factors in scoring. Different satistical concepts, e.g. reliability, validity, difficulty and discrimination indices were developed. Interpretation of examination results also underwent a change. Methods of scaling were evolved for interpreting marks. Test standardisation and establishment of norms completely revolutionised the examination movement. A new discipline known as "Psychometry" (by the British) or "Psychometrics" (by the Americans) was developed, which came to be regarded, as Harper (1960, p. 1) puts it, "both a science (based on rigorous statistical and research methods) and an art (characterised, at its best, by a high degree of technical skill, creative ingenuity, and insights into human behaviour.)"

In the course of research the same experiments were often repeated at different places. This was probably because of a general feeling that

since teaching and learning conditions, which vary from place to place, have much bearing on examinations, research findings of one place may not be universally valid The following statement of the Examination Committee (1962, p 5) alludes to the above feeling

> Methods of testing should be related to practices of teaching and conditions of study, and should be capable of adjustment and adaptation to particular situations and needs

The same feeling is voiced by Dave (1966, p 9)

> Successive commissions, committees and study groups on education have pointed out the evils of this examination system The remedies to all these evils cannot be found in the researches done elsewhere in the world, and only careful studies and investigations can offer solutions to problems vexing the people concerned with education.

Definition of Terms Used*

The two basic concepts of reliability and validity in examinations are used throughout this chapter and are understood to mean the following

RELIABILITY

This "refers to consistency throughout a series of measurements" (Cronbach, 1960, p 126) In other words, "The reliability of any set of measurements is logically defined as the proportion of their variance that is true variance" (Guilford, 1956, p 436)

The problem of reliability of the essay tests may be divided into three categories

Reader Reliability

The "correlation between the marks of two readers is known as 'the reader reliability' of the examination" (Gulliksen, 1950, p 211)

Content Reliability

"The reliability of an essay test corrected for attenuation due to the inaccuracy of reading has been termed the content reliability of the essay test". (Gulliksen, 1950, p 214)

Total Reliability

This is a term coined by Harper (1966 a, p 12) and defined by him as "the correlation between parallel examinations marked by parallel examiners That is, it is the correlation between examination 1 marked

* In this chapter, terms have been used which have currency in the literature In the next chapter we will re-define some of these more critically, for the purposes of the present research study

by examiner A and examination 2 marked by examiner B This is the reliability coefficient with which we should be most concerned in comparing the reliabilities of essay and objective examinations, as well as in practical decisions of the degree of trust to be placed in an examination mark"

VALIDITY

"A test is valid when it measures well what it is supposed to measure" (Micheals and Karnes, 1950, p 104) To be valid a test must be reliable But the reverse is not true A test may be highly reliable, but it may be invalid for the purpose it is supposed to serve

The following terms have been used interchangeably in the study

"reader" and "examiner"

"essay test", "essay examination", "essay-type test" and "essaytype examination'

"objective test", "objective examination" "objective type test' and "objective type examination"

"score reliability , "test reliability' and "total reliability' "grading reliability", "marking reliability', "examiner reliability" and "reader reliability'

Summary of Early Studies

A comprehensive and scholarly review of the studies on examinations done prior to 1958 has been made by Basumallik (1959) Unfortunately, the review is marred by a few points, some minor and some major, which may mislead the lay reader Without in any way detracting from the over-all value of the review, it seems wise at this point to call attention to these points

While introducing the concept of reliability (p 128) Basumallik states, "In a general sense reliability means the accuracy with which a measuring device measures *whatever it is expected to measure*" (Italics are the present researcher's) The definition, as it stands, best fits the definition of validity The italicised portion should have been replaced by some thing such as "whatever it does measure" Many researchers would disagree with Basumallik's statement (p 149) where he says

> It is needless to reiterate that in comparing essay and objective test we should not consider reader reliability for obvious reasons Comparison should legitimately be made between test reliabilities of the two types of examination The test reliability, for essay examination, should be the content reliability or the reliability estimated by the methods of test re test, analysis of variance, etc It is desirable to compare coefficients of a similar nature

On comparability of essay and objective tests, Gulliksen (1950, p 212) observes "If two parallel essay examinations are matched just as successfully with respect to content as are two parallel objective examinations, the correlation between the two parallel essay forms will practically always be lower than between the two objective forms, owing to the fact that the unreliability of reading will still further lower the correlation between the two essay forms" Thus, as Harper (1966 a) holds, for all practical purposes only total reliability, *and not content reliability*, of the essay test should be compared with reliability of the objective test Since reader reliability of the objective tests is 1 00 (i e the objective examinations completely eliminate this source of unreliability), ignoring it in the essay examinations presents an unrealistic picture

Test reliability is a function of test length This Basumallik has also noted in his review (p 150) Obviously, a 24-hour or so essay test (as the Chinese seemed to have) would be more reliable than a half hour objective test Hence for comparing reliabilities of the essay and objective tests the testing time should also be equated But while giving comparative figures of the essay and objective test reliabilities (p 150) Basumallik does not record testing times Explaining the comparative figures of the essay and objective test reliabilities, he (p 151) states, "The median reliability coefficients, however, are 85 for essays and 93 for objective tests respectively" The direct comparison of reliability figures implies that testing time for all the tests whose reliabilities are under comparison is the same But his further discussion indicates that testing time varied for different tests It may be pointed out that direct comparison of the reliabilities of tests of unequal lengths is technically objectionable and is misleading to laymen

Basumallik further states (p 151) that 'the difference between the two (85 and 93) is only 8, in favour of objective tests' (Obviously 8 is a printing mistake for 08) Similarly, at another place on the same page he states that "These reliabilities range from 80 to 96, with a median of 90 which is only 05 greater than the median essay reliability stated above" (i e 85) In explaining differences between reliability figures he ignores a vital point that reliability figures do not represent a linear scale It is z co-efficients whose scale is linear Thus he should have converted the reliability figures to z's, obtained the difference between the two z's again converted the obtained difference of the two z's to r, if he was interested in finding the difference between the two reliability figures The present researcher doubts that even this approach, though statistically sound, would be able to give any idea about the relative merits of the two tests The best and probably most easy to understand approach for such comparison seems the application of the Spearman-Brown Prophecy formula This formula tells us that a test of reliability of 85

has to be lengthened 2.34 times to reach the reliability of .93. In other words, the difference between the reliability figures of .93 and .85 is not nominal, as Basumallik's inference seems to sound. This point should have been made more clear in his review to avoid any misconception in the minds of non-technical people.

Despite the above-noted minor negative criticisms it may again be emphasised that Basumallik's is an excellent review and the quality of his work may not be looked down upon for some minor lapses here and there. The following conclusions are drawn from his review.

One must say that reliability of a usual essay-type examination—reader as well as content—is not as high as that of an objective test. This is true even of subjects like mathematics and chemistry. An objective test, even when imperfectly constructed, is not likely to have reliability lower than .70, whereas the usual essay test will have still lower reliability. It may be a problem for further experimentation to find out how best an essay test can be improved. Under certain conditions, e.g. team impressionistic marking, however, an essay examination can be so improved as to stand fair comparison with an objective test in point of test reliability. But these conditions are not always feasible to employ in actual situations involving a huge number of examinees and requiring quick decision. More often we have to depend upon a single marker in a real situation, and the marking of a single examiner is found to be considerably erroneous. On the whole, the research on essay reliability, done so far, presents incomplete and controversial evidence on many issues. Intensive research is needed to see how far it is possible to improve the reliability of an essay.

Basumallik has reviewed the earlier studies so well that it did not seem necessary to repeat this work. The present researcher, therefore, confined himself to reviewing later studies as well as some important studies omitted by Basumallik.

Examination Research Abroad

I Conversion of Essay Items to Objective Items

Some foreign and Indian experts (Vernon, 1940; Wrightstone, Justman and Robbins, 1956; Thorndike and Hagen, 1961; Examination Committee, U.G.C., India, 1962; Standards Committee, 1965) hold that the essay-type tests measure those larger outcomes of education which may not be measured by other paper-and-pencil-type tests. Others (Sims, 1948; Stalnaker, 1951; Noll, 1957) hold that this claim of the essay-type tests is generally accepted without proof, evidence, or supporting logic.

Implicit in the claim of essay type tests is that objective tests do not and cannot measure the larger outcomes of education Some experts (Ballard, 1923, Diederich, 1957, Leo Nedelsky, 1957, Engelhart, 1957, Ryans, 1958, Frederiksen, 1960, Kelly, 1963, Harper, 1963 (b), Hubbard, 1963, Hill, 1964 a) hold that most of the educational outcomes measured by the essay tests can be measured by good objective tests Ebel (1953) and Anastasi (1966) hold that the impression that the objective tests can measure only trivial educational objectives is probably due to the fact that such objective tests can easily be prepared by unskilled writers

However, some studies have tried to find out empirically as to what extent the material covered by the essay tests can be made objective

Sims (quoted in Woods, 1953) showed that out of 243 essay questions in elementary education examinations, 103 were simple recall 97 short answer and only 43 discussion type questions In secondary education examinations, out of 215 questions, 54 were simple recall, 63 short answer and 98 discussion type questions He concluded from his study that only "30 5 per cent of the questions (were) in reality subjective questions and better than two-thirds of the grand total studied could have been converted to objective type tests'

Orleans and Sealey (quoted in Woods, 1953) while not making a statistical study of the problem, did call attention to it by giving an essay test and then converting it to an objective type test to show the thesis of objectivity in essay questions

Woods (1953) found that in one hundred separate essay tests studied, there were a total of 1 456 answer items expected as shown by the teachers' suggested answers Of that number, 1,442 were capable of conversion to objective type tests The study suggests that if the essay type questions which can be converted to the objective type tests are handled thus, there would be much saving of teacher pupil time, and students would be better able to understand the fairness of the grading The study also suggests that teachers use the essay tests where objective tests would have done the job better, and teachers frequently do not accomplish the expressed goal of the essay tests of measuring pupil s thinking organisation and expression due to improperly worded questions

Moore (1954), giving a concrete example has shown how one essay question (for graduate medicine) can be covered by seventeen objective items Assuming that this example is average, he concluded that a five-question essay examination may be covered in eighty five objective items The usual objective examination of two-to three hour duration contains two to three hundred items, which could correspond to twelve to eighteen essay questions Thus, the objective examinations can cover a broader spectrum of knowledge than the essay examination in the same time period

II Reliability and Validity of the Essay Tests

Several investigators have explored the reliability and validity of essay type tests

Wiseman (1956) had 30-minute compositions written by a group of 173 students and graded by four examiners This exercise was repeated with another title four months later, and the essays were graded by the same four examiners Correlation between the aggregate marks on the first and second essays was 89, corrected for the range of ability in the sample, it was calculated to be 92 This means that the total reliability of the essay tests is high for the pooled marks of four readers

Wiseman (1956), in a validity study on a different sample (N=117), correlated entry test scores including an essay with the following five criteria

1 Total School Certificate results
2 School Certificate English Language
3 School Certificate English Literature
4 Teachers' over all estimate
5 Teachers' estimate of written English

Multiple *r*'s calculated for the whole battery, and for the battery without the essay mark included, revealed that the addition of the English essay increased the size of the coefficient in the case of all the five criteria However, it may be added that the design of this experiment is open to criticism on the following grounds

Selection of the sample was biased in favour of students regular in their studies It is also probable that school certificate standards differed significantly from school to school, because all the schools selected for the study were not of the same quality Hence the inference based on the correlations with school certificate may be misleading Thus, the findings of this study may only be considered tentative pending corroboration by further evidence

Penfold (1956) explored variability in marking essay tests Five-minute essay written by 16 000 candidates were marked by impression by sixteen examiners, each examiner marking one thousand scripts After an interval of time (the exact period is not reported) 165 scripts selected at random from the whole batch were all re marked by fifteen of the same examiners, the sixteenth not being available The results indicated that "there was significantly high variation between different examiners' standards of marking and also between the marks of the same examiner on the two different occasions of marking"

From the above results Penfold anticipated that in all probability the best chance of improving consistency would be to devise an analytic system of marking longer essays He, therefore, studied the analytic

marking of 20-minute essays Sixteen examiners were involved in marking including the chief examiner At a standardising meeting the examiners discussed the details of marking To be quite sure of the conformity in marking each examiner marked a photostat set of 25 scripts judged by the chief examiner to cover the whole mark range Finally after a prolonged period of marking to this scheme the extent to which the errors had been reduced was checked by an analysis of marking of twenty more photostatted essays The results suggested significantly high variations in marking between the examiners in spite of the extensive training and experience

It may be observed that no examining body can afford to give that much intensive training to its examiners as was given in this experiment This is another evidence to prove that variations between the markings of examiners is an inherent and incorrigible weakness of the essay type tests

Pidgeon and Yates (1957) found test re test reliability for the essay tests to be 77 if the same examiner re marked and 72 if a different examiner re marked In a follow up study of 473 students who had completed their secondary school course, they found that the old type (i e essay type) English examination provides significantly less satisfactory forecast of subsequent success in secondary schools than is obtained from the results of an objective test

Hudson (1960) found that (1) neither the Cambridge F R S s nor the Cambridge D Sc s had better degree classes than those of their respective control groups and that (2) although Cambridge F R S s were roughly three times as numerous as those from Oxford they were far less likely to have a first class degree He concluded that if our result is accepted as approximately accurate it would appear that class in Tripos bears no direct relation to subsequent eminence in scientific research as assessed by either of our two criteria

Valin (1961) found total reliability of the essay tests by correlating marks on two tests to be 12 for Arabic Philosophy, and 52 for Mathematics Valin has not reported testing time As such it is not clear whether the very low reliability of Arabic tests was due to short lengths of the tests or due to subjectivity in the essay type tests

French (1961) got 300 short essays written as home work by college freshmen marked independently by ten English teachers nine social scientists, eight natural scientists ten writers or editors nine lawyers and seven business executives on nominal compensation The examiners were asked to use their own judgement in marking and sort the papers into nine piles in order of merit The average correlation between all the examiners was 31 Out of the 300 essays 101 received all nine different grades and no paper received less than five different grades. The average correla

tion between English teachers was 41 The higher correlation among English teachers was attributed to their less erroneous markings He concluded that nothing at all is being measured very well by the essays

Many teachers feel that reader reliability may be improved if essays are marked on generally accepted criteria Fostvedt (1965) explored the criteria for evaluating high school English composition Using generally accepted criteria, thirty experts gave numerical ratings to twenty themes He concluded that perhaps the only conclusion justified by the study was that, although teachers of English composition might feel that criteria are important in evaluating these, there was no evidence of consistency in the employment of such criteria

Lehtovara (1966) in a study on a sample of 1330 students concluded that the validity of the essay examination could be improved by the modern test theoretic methods The reliability of the grading of test papers was found to be 85 This is one of the very few western studies where reader reliability of essay examinations was found to be above 8 It may be recalled that the grading reliability of 85, which looks very high, is far below the reader reliability for the objective type tests For an objective test scored carefully with an agreed key, the reader reliability is 1 00 Even for careless scoring, the reader reliability of the objective tests is far higher than 85 Ebel (1965 p 312) remarks "Even if over half of the answer sheets in a set of thirty-five were scored wrongly, with errors like those shown below, the coefficient of scoring reliability would still be closer to 99 than to 98"

						Total
Size of Scoring Error	4	3	2	1	0	
Number of Errors	2	2	5	11	15	35

It may be stated that a test of reliability of 85 has to be lengthened approximately nine times to reach the reliability of 98

Myers (1966) had essays of over 80,000 students double marked In a period of five days, 145 readers gave each of the essays two independent readings on a four-point scale The average reliability of the judgements calculated by analysis of variance method among all readers and across all papers ranged from .26 to 49 over a five day period

One of the arguments constantly brought up in favour of the use of the essay examinations is that they purport to measure creativity However, Page (1967) in a study of a sample of essays written by secondary school students judged by independent examiners found the measurement of creativity on essay tests to be the least reliable attribute rated by human judgement, and mechanics the most reliable

Coffman (1966) investigated the validity of the essay type tests to predict writing ability The students who were the subjects of the research were in classes XI and XII in 24 American secondary schools Each student took six objective and two semi objective tests and wrote five essays of different lengths Each student wrote on five different topics and each essay was marked by five trained examiners The sum of these twenty-five marks was the student s criterion score The reader reliability of one essay read once, that is, the correlation between two independent readings of one essay ranged from 11 to 34, the average being 26 which is a very low figure The validity coefficient of essay tests ranged from 33 to 55 He concluded that "in order to obtain validity coefficients comparable to those obtained from an one hour objective test using only essay tests, it would be necessary to assign at least two topics to each student and have each read by five different readers, or to assign three topics and have each read by three different readers" Vernon (1940) makes exactly the same recommendation for attaining a high reliability for the essay-type test, that is, to have a minimum of three essays from each pupil, each marked by three markers A further analysis of the data of Coffman et al (quoted in Hill, 1967 e) showed the following major findings

1 "'The reliability of essay scores is primarily a function of the number of different essays and the number of different readings included' These factors are more important than the length of the essays, and they have more value than analytical reading"

2 "Validity is lost when optional topics are given" In a situation where choice is allowed, "a student's rating might depend more on which topic he chose than on how well he wrote"

3 "'When objective questions specifically designed to measure writing skill are evaluated against a reliable criterion of writing skill, they prove to be highly valid'"

4 "'The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions in combination with objective questions' A combination of two objective tests and one essay, requiring a total testing time of one hour, correlated most highly with the criterion However, the validity added by a twenty minute essay is small, and it requires that the essay be marked by at least three examiners 'It is doubtful that the slight increase in validity alone can justify the increased cost'"

Bracht and Hopkins (1968) found reliability coefficients of two 20-minute objective tests, calculated by K R -20, to be 40 and 50 The reliability coefficients of three essay tests of the same time length, calculated by the analysis of variance method (that is content reliability), was found to be 58, 26 and 69 As already observed, the content

reliability of the objective test is comparable to the total reliability of the essay-type test. If reader reliability were taken into account, the essay test reliabilities reported would be reduced considerably, while the objective test reliabilities would remain unchanged. Thus the essay and objective test reliabilities as reported here should not be considered as directly comparable. The other major findings of this study are as follows:

Cumulative grade point average correlated very significantly with performance on both types of tests, but the correlation with the objective tests was significantly higher. Writing ability was significantly related to both the essay and objective tests, the difference between the correlations being not significant. The findings of this study do not support the common supposition that the objective and essay tests measure different abilities. In addition, there is clear evidence that the objective test is a more reliable measure of academic achievement.

Klein and Hart (1968) investigated the chance and systematic factors affecting essay grades by getting the scripts of over 1500 students of Law marked by seventeen professors. They found that students tended to be consistent in their performance in different essay questions used in the law school. The agreement in grading among the professors was .76.

III. Conclusions Suggested by the Foreign Studies

1. Most of the material covered by essay tests may be converted to objective tests. Much more content area may be covered in testing if essay tests are converted to objective tests.
2. For short essays, reliability is very low.
3. There is significantly high variation between essay markings of different examiners, or of the same examiner on two different occasions. Intensive training for standardising markings seemed to have little effect in reducing the variations in markings of different examiners.
4. Creativity, which is supposed to be the major quality assessed by the eassy test, is least reliably graded.
5. In essay examinations in order to get a validity coefficient comparable with that of a one-hour objective test, it is necessary that at least two topics each marked by five independent examiners or three topics each marked by three independent examiners be obtained.
6. The essay and objective tests do not necessarily measure different abilities. On the other hand, the evidence shows that objective tests are significantly better measures of academic achievement
7. Validity is lost when optional questions are provided, and a

student s score may depend more on what topic he wrote than on how well he wrote

8 The predictive validity of essay tests in forecasting success in succeeding examinations is significantly less than that of objective tests

9 Essay examination results are not valid measures for predicting eminence in scientific research

Examination Research in India

Many areas in examinations have been covered by research studies in India These studies can be broadly classified in two sub heads Functional aspects of examinations and efficiency aspects of exam nations

1 Functional Aspects

Studies on the functional aspects of examinations have produced what is essentially a survey of the statistics about different examinations in India A Mysore University survey (1965) showed that pass percentages for Pre University examinations vary from university to university in India In 1965 the pass percentage for the Pre University examination for the Mysore University was 30 whereas it was 55 for the Bombay University The pass percentage for the same examination even for the same university varies from year to year The pass percentage for the Intermediate examination varied from 26 in 1935 to 48 in 1945 in the Mysore University Mukerjee (1954) in a study done in West Bengal found that the pass percentage for the Matriculation examination varied little from 1912 to 1945 He inferred that pass percentages remain almost static from year to year Dave and Patel (1966) DEPSE (1962 a and Undated a b) and the S S C Board Maharashtra (1963 a) found that the pass percentages of regular students are markedly higher than that of the private candidates The S S C Board Maharashtra (1960) found that there was better pass performance among urban students than among their rural colleagues They found that at the age of 19 20 years girl students do better than boy students while at the age of 16 19 years boy students do better than girl students It was also noticed that the mean percentage of successful candidates rapidly declines with the advance of age above 16 years This tendency was found both for boys and girls The regional languages were found to be more efficient as the medium of instruction than the English language in enabling the students to score higher marks in Science History and Geography

Chitrakara (1961) DEPSE (undated a) Deshmukh (undated) Bhanot (1961) Deshmukh and Kamat (undated) The M S University Baroda (1964) the Mysore University (1965) Pillai (1965) and the

S S C Board, Maharashtra (1960, 1963 a) studied the wastage and stagnation among students The wastage and stagnation were found for each class to be between 40% to 70% The general causes of the wastage and stagnation as inferred by these studies are poverty, lack of proper atmosphere at home and school, want of equipment and qualified staff in schools Lele et al (1966) analysed the failure rates of the Preparatory Commerce examination of the M S University, Baroda for the years from 1950 51 to 1964 65 (The Preparatory Commerce examination is a one year course where students are admitted after passing the S S C Examination) They found that there was a significant upward trend in the failure rate in the Preparatory Commerce examination Lele et al inferred that assuming that the quality of the S S C examination, as evinced by the S S C examination marks, remains the same, the Preparatory Commerce examinations have become more difficult

Mitra (1958) and Misra (1968 a) factor analysed examination marks of students to find out common factors underlying different academic subjects Mitra's study does not provide any conclusive evidence, probably because the number of cases in the sample was very small Misra (1968 a) found three factors in the Matriculation examination of the Gauhati University The first factor, which was dominant, was named "verbal' The second and third factors were almost of equal order, and were named "problem solving" and "memorisation" Contrary to his expectation, Misra did not find a numerical factor He stated as the probable cause that the condition of having at least two variates for each ability was not satisfied in his experiment in the case of the numerical factor Lele et al (1964) factor analysed the S S C and P Sc examination marks, the two being successive examinations, for the years 1957-58, 1958-59 and 1959-60 They found two common factors, namely, "verbal" and "numerical"

Roy (1969) found that the failure rate in Mathematics in H S L C examination is on the increase It has shot up from 45 7% in 1966 to 76 9% in 1968 He attributes the reason of failure to the poor quality of students appearing in the H S L C Mathematics examination

The studies on the functional aspects of examinations provide useful information but their number is very few On some points the findings are contradictory, which suggest the need of further investigation Moreover, the value of such studies is limited as regards the assessment of the effectiveness of examinations as a tool of measurement

II Efficiency Aspects

To be an efficient tool of measurement, an examination should be

valid precise and economical Studies on the efficiency aspects of examinations may be classified in the following sub heads variation in markings marking errors reliability of marks and validity of marks

I VARIATION IN MARKING

(a) *Variation for the same set of scripts*

There are several studies on the variability of marking the same set of scripts by different examiners or by the same examiners on different occasions

Mukerjee (1961) had two essays one in Assamese and the other in English each marked independently by five competent examiners The five examiners for the Assamese scripts awarded marks ranging from 14 to 29 The five examiners for the English scripts awarded marks ranging from 6 to 22. The two scripts were marked by the same examiners after a week The highest difference in marks given by *the same examiner to the same scripts* on the two occasions was 12 (i e 31-19) for Assamese, and 11 (i e 22 11) for English The mean marks awarded to the Assamese scripts varied from 26 4 to 18 6 and for the English script from 15 6 to 8 4 on the two occasions As is evident from the study not only did the examiners differ from each other, they even differed from themselves on the two occasions—i e after a lapse of seven days

George (1964) studied the difference in examiners markings in the Pre University Test of the Kerala University by selecting a random sample of 40 scripts each of English I English II Mathematics Physics Chemistry, Botany and Zoology These scripts were marked independently by two separate external examiners of the University Except for English I and Zoology in all other subjects the differences between the markings of the external examiners were found significant at the 1% level In Zoology the difference was found significant at the 5% level

George (1964) in another experiment had a large number of scripts marked by two independent University examiners in the above noted subjects of the P U students of the Kerala University He found that the percentage of disagreement in classifying students in different categories that is I II III and Fail ranged from 22 6 in Mathematics to 52 2 in Botany In Physics one student who was *failed* by one examiner was given a *first class* by the other examiner In Botany there were two such students In Mathematics one student who was given a *fail* and twelve students who were given a *third class* by one examiner were given a *first class* by the other examiner

It may be observed that the scripts used in both the experiments of George (1964) seem to have contained crosses and ticks given by the

original examiners. The crosses and ticks might have invited the attention of the other examiners to the points marked by the original examiner. Thus, the obtained disagreement in the markings should be considered as the minimum obtainable in a real examination situation.

Taylor, Tluanga and Misra (1966 b) had answer scripts of 100 examinees of B.A. Part I English examination of the Gauhati University cyclostyled and examined independently by nineteen examiners. The mean marked from 19.2 to 35.0 (out of 100 marks) and pass percentages ranged from 7 to 62 for the nineteen examiners. In some cases the disagreement in awarding marks was extreme. For example, it was found that a student who got a *3rd rank* from one examiner was given a *98th rank* by another. These scripts were originally marked by the University examiners. Out of the 100 students, the number of original failures who passed and the original passes who failed in the second marking ranged from 20 to 37 for the nineteen examiners. It may be observed that in the experiment the copies of scripts were not reproduced in the students' own handwriting. They were typed. Thus, dis-agreement in marking due to the quality of handwriting was eliminated. As such, the disagreement as reported in the study should be considered as the minimum expected in a real situation. Harper (1967 a) had facsimile copies (by Xerox process) of 10 scripts of History examination of class X of a Board of Secondary Education marked by 90 experienced examiners of the same Board. It was found that the only student considered fit for the award of a star (i.e. above 75% marks) was also failed by several examiners The script which was given first or second rank by almost all the examiners got from a high first class to a bare pass mark. The variation in pass percentages ranged from 10 to 80. This is the only study in India where variation in markings was explored in an almost real examination situation, i e. the scripts were reproduced in students' own hand-writing, these were sent to examiners during the time when they usually examine the Board scripts, and the examiners were paid remuneration at the Board's rate. Thus, the findings of this study may be considered as representative of variability in markings in real examinations.

Harper et al. (1967) had scripts in Hindi, History, Geometry and Biology of Class X final examination of a Secondary Board re-marked, some by the same examiner and some by another examiner. The highest variation in marks was found in Geometry (supposedly an "objective" paper). The same examiner, re-marking his own scripts differed from himself by 42% marks. The highest *mean* difference was found in Biology (another supposedly "objective" paper), a mean difference of 9.8 for ten examiners each re-marking 50 of their own scripts They found that the variability in marking remains almost of the same magnitude

whether an examiner re-examines the scripts previously marked by himself or marked by another examiner

Harper et al (1967) analysed the variability in marking in a different way also In the above-noted experiment the markings of the two examiners were equally erroneous They, therefore, selected one examiner who was considered by the Board to be highly experienced in marking, that is, a "standard' examiner The scripts marked by the "standard" examiner were re-marked by other experienced examiners of the Board The mean variations in the two markings were found to range between 14% to 32% of marks in History, 12% to 24% in Hindi 12% to 16% in Biology, and 18% to 28% in Geometry

Many teachers feel that marking is objective in Science and Mathematics papers The experiments of George (1964) and Harper et al (1967) show that in essay type tests subjectivity of markings in the Science and Mathematics papers is no less, if not more than that in the Arts Papers

Taylor (1962 e) had scripts of 45 students in English (two papers) Economics, History, Logic and Mathematics independently marked by two examiners The highest difference between means was found in Logic where the mean mark of one examiner was 55 8, and of the other examiner was 46 0 The lowest difference between means was found in History, where the mean marks varied from 40 4 to 39 5 The highest difference in the standard deviations was in Logic, where the s d 's varied from 18 0 to 10 8 He concluded that "the experiment makes it clear that an examination mark has neither the sanctity nor the precision which is usually attached to it" He recommended the use of scaling to minimise the effects of errors in marking Some people suggest that to improve the reliability in marking essay type tests, fewer number of scripts should be given to examiners This experiment makes it clear that even for a very few scripts the differences in means and s d s in the markings of two examiners are alarmingly large

(b) Variations for equivalent sets of scripts

Taylor did several experiments on the variation in markings for statistically equivalent sets of scripts He observed that, "By allotting roll numbers at random, or otherwise mixing the candidates one can ensure that the sub-groups are statistically equivalent"

He used this method of making sets of scripts statistically equivalent in the following three studies, so that any differences in the mark distributions can be attributed to the examiners and not to the candidates

Taylor (1962 b) found that in an English paper of college students one examiner produced a distribution which Taylor describes as 51

(5, 3) (where 51 is the mean, 5 the difference of the upper quartile from the mean, and 3 the difference of the lower quartile from the mean) and the other one for the equivalent set of scripts produced the distribution 31 (3, 4) With one exception *all* the marks given by the second examiner were lower than *any* marks given by the first examiner

Taylor (1962 b) had a first year college paper in Biology, involving 210 candidates marked by two pairs of examiners, A and B taking 110 scripts, C and D taking the remaining 100 scripts The two sets of scripts had been made statistically equivalent by allotting random roll numbers The marks of each pair of examiners were averaged The median mark was 66 5 for A and B, and 42 0 for C and D The true means were respectively 64 9 and 40 4 Taylor concludes

> The difference in the average marks (whether one used the median or the mean) is thus 24-1/2 With A and B, 88 per cent of the candidates got 55 marks or more, with C and D on the other hand 95 per cent of the candidates got less than 55 marks The difference is even more remarkable in that each total mark is found by combining the separate marks of two examiners a procedure which would tend in general to smooth out differences in the standards of marking The examiners were very unwilling to admit even the possibility of such large differences until the analysis was put before them

Taylor (1963 a) found the pass percentages of the equivalent scripts marked by different examiners varying from 11% to 96% in English 27% to 77% in Mathematics, and 15% to 70% in Science, in the Matriculation examination of the Gauhati University

2 *MISCLASSIFICATION OF STUDENTS DUE TO MARKING ERRORS*

Some people feel that classification in divisions is reliable though the marks may not be so Some Indian studies have explored this point

Taylor (1962 c) studied a hypothetical group of one thousand students with an average mark of 45 and a standard deviation of 15 He assumed that a pass was secured by a mark of 30, second class by a mark of 48, and a first class by a mark of 60 He further assumed the standard error of marking of 5 marks He found that about 25% of the students would be wrongly classified in I, II, III and Fail categories in the above-noted case He also found in this case that 9% of the students who really deserve a pass would fail on a paper by accident On this assumption, he calculated the probability of passing a student in a ten paper examination when the student has the same ability in all the papers and would pass the whole examination on the basis of his

TABLE 5.1

TABLE SHOWING RELIABILITY FIGURES OBTAINED IN INDIAN STUDIES

	Lang-uage	Hindi I	Hindi II	Hindi I+II	Bengali I	Bengali II	Bengali I+II	Classical Sanskrit Language	English I	English II	English III	English I+II+III
Content Reliability Gayen et al.		.80 (X)	.68 (X)	.84 (X)	.74 (X)	.65 (X)	.70 (X)	.85 (X)	.7 (X)	.7 (X)	.6 (X)	.85 (X)
Harper et al.		.45-.77 (X)										
Misra									.76 (X)	.63 (X)	.63 (X)	
Misra									.8 (P.U.)			
Reader Reliability Taylor et al.									.75 (B.A.I)			
Harper												
Harper et al.		.79 (X)										
Total Reliability Gayen et al.					.39 (X)	51 (X)	.55 (X)		.61 (X)	.63 (X)	.44 (X)	.73 (X)
S S C. Board Maharashtra	·49-.51 (X)							68- 69 (X)	.69 (X)			
Harper et al.		.58 (X)										
Misra									.64 (P.U.)			

TABLE 5.1 (*Contd*)

TABLE SHOWING RELIABILITY FIGURES OBTAINED IN INDIAN STUDIES

	Social Studies	History	Geo-graphy	Civics	Maths	Arith-metic	Algebra & Geom.	Gen. Science	Science	Bio-logy	Phy. I	Phy. II	Phy. I+II	Chem. I	Chem. II	Chem. I+II	Aggre-gate Total
Content Reliability																	
Gayen et al					63 (X)						80 (X)	80 (X)	.83 (X)	.75 (X)	.75 (X)	.80 (X)	
Harper et al.		.21-.76 (X)			68-.90 (X)					.41-.64 (X)							
Misra		.60 (X)	49 (X)		.50 (X)												
Misra																	
Reader Reliability																	
Taylor et al.																	
Harper		.83 (X)															
Harper et al.		.77 (X)			.96 (X)					80 (X)							
Total Reliability																	
Gayen et al.		43 (X)	35 (X)		.60 (X)												
S S C. Board, Maharashtra	.46-.52 (X)	.52-.53 (X)	.40-.54 (X)	33-.53 (X)	68-.79 (X)	.61-.70 (X)	.71-.78 (X)	.54-.62 (X)	.57-.60 (X)								.76-.80 (X)
Harper et al.		.35 (X)			.72 (X)					.50 (X)							
Misra																	

true mark The probability of passing simultaneously in a ten paper examinations was (for such a student) 39

Harper (1963 b) showed that due to the provision of choice, students of the same ability may get different marks Taking an example where two students of the same ability appear for a question paper in which five out of nine questions are to be answered, he demonstrated how one student may get 75 marks (a distinction) and the other may get 65 marks (a first class) out of 100

Harper et al (1967) got 1,000 scripts of class X examination in each of Hindi, History Geometry and Biology double-marked They found the expectancy of getting the same division if marked again as follows In History 9% for I division, 42% for II division 64% for III division and 63% for fail In Hindi 36° for I division 53% for II division, 66% for III division, and 67% for fail In Biology 39% for I division, 62% for II division, 64% for III division and 63% for fail In Geometry 84% for distinction, 66% for I division 65% for II division, 61% for III division and 90% for fail

George (1964) got 3,416 scripts of English Mathematics Physics Chemistry, Botany and Zoology of P U students of the Kerala University re marked He found that 'results of nearly 30° of the students would have been altered in terms of pass divisions or failure either to their advantage or disadvantage by re valuation ' He came to the conclusion that ' the present system of evaluation produces considerable degree of error

3 *RELIABILITY*

Studies on reliability have been grouped under three heads Content Reliability, Reader Reliability and Total Reliability

(a) *Content Reliability*

As defined earlier, the reliability of an essay type test corrected for attenuation due to the inaccuracy of reading has been termed the content reliability of the essay test

There are studies by Gayen et al on the content reliabilities of certain papers of Class X examination of the Board of Secondary Education, West Bengal Gayen et al do not make any differentiation between various types of essay test reliability They have used only one term "reliability" for the various types of reliability reported by them The studies reported under the head ' Content Reliability have been grouped by the present researcher, as they seemed to him to fit the conditions of the content reliability

Gayen et al studied the content reliabilities for several papers of

class X examination of the Board of Secondary Education, West Bengal

In the first study (1961) the content reliability was found to be 63 for Mathematics In the second study (1962) it was found to be 7 for English I and II each, 6 for English III, and 85 for all the three papers taken together In the third study (1966) it was found to be 85 for Sanskrit, 80 for Hindi I, 68 for Hindi II, 84 for Hindi I and II taken together, 74 for Bengali I, 65 for Bengali II, and .70 for Bengali I and II taken together In the fourth study (1967) it was found to be 80 for Physics I and II each, 83 for physics I and II taken together, 75 for Chemistry I and II each, and 86 for Chemistry I and II taken together In the above studies each paper, except English II which was of two hours, was of three hours

It may be observed that in India the above-noted studies done by Gayen et al are the pioneering studies analysing a Board's results by modern psychometric methods The strengths of these studies are many There is no study in India on public examinations where the sample is so representative of the population as in these studies These are the only studies in India which have tried to analyse every aspect of examinations using a wide variety of statistical tools of analysis Of course, from a pioneering study, one cannot expect to get answers to each and every problem in the field It is also natural that even such elaborate studies may have some weak points Without in any way detracting from the overall value of the studies, it appears desirable to invite attention to the following points

Their definition of difficulty index is novel i e different from generally accepted definition of difficulty index Unless one knows the special meaning given to this term by them—and it is very unlikely that a technical man would consult their definition due to his confidence in the knowledge of the term—he will certainly draw wrong conclusions from their figures of difficulty index Commenting on the generally accepted definition of difficulty index, Ebel (1965, p 359) rightly states that "This inverse relation between what is ordinarily meant by difficulty and the index used to measure it is illogical, but the convention has become so firmly established that to attempt a change would be only to contribute to confusion"

While discussing the concept of difficulty index, Gayen et al stated that the difficulty of an item depends upon the ability of the group who answered the item But in the calculation of the difficulty index of an item, they seem to have ignored the ability level of the students answering the particular item

In splitting the Mathematics paper in two halves they almost ignored the matching of course content, and were mainly guided by the consideration of having the other half which could best correspond

with the marks awarded in the first half This approach may lead to spurious reliability In the application of K R -20, they have treated the total of marks awarded on the several sub groups of a question as the item score for that question This approach would also lead to spurious reliability However, they have amply made it clear in a true researcher's spirit that these were the only approaches which they thought they could possibly adopt, though they knew that their data did not satisfy all the conditions needed for such approaches

Despite these minor criticisms, it may be emphasised that Gayen et al have set the tone for analysis and interpretation of many problems of essay-type examinations, and their endeavour deserves all commendation To quote Shakespeare

Frail creatures are we all, To be the best
Is but the fewest faults to have

Misra (1968 a) investigated content reliabilities of English I, II and III, History, Geography and Mathematics of the Matriculation examination, 1963 of the Gauhati University by factor analysing the marks obtained by 1,241 students selected from the examination on the basis of a random sampling He found the content reliability (obtained h²) to be 76 for English I, 63 for English II, 63 for English III, 60 for History, 49 for Geography, and 50 for Mathematics

Misra (1968 b) in another study found the content reliability of English I of three-hour duration of the Pre-University examination 1966 of the Gauhati University by analysis of variance, and split half methods to be around 8

Harper et al (1967) investigated the content reliability of Hindi, History, Biology and Mathematics of class X examination of a secondary board of education They found the content reliabilities for various examiners in Hindi ranging from 45 to 77, for History ranging from 21 to 76, for Biology ranging from 41 to 64, and for Mathematics from 68 to .90 They attribute the variability in the content reliability of the same question paper for various examiners as due to the fact that different examiners may emphasise different aspects of the answer, e g style, factual accuracy, organisation, creative thinking, etc Since the proportion of each of these in the content of answer scripts may differ, the emphasis on different aspects is likely to produce different content reliabilities for different examiners

It may be noted that Gayen et al (1961, 1962 and 1966), Misra (1968 a and 1968 b) and Harper et al (1967) found approximately the same reliability figures for different papers, except Mathematics. For the Mathematics paper Harper et al (1967) report high reliabilities. i e. 68 to 90 (median 77), and Gayen et al (1961) and Misra (1968 a) report low reliabilities, i e 63 and 50 respectively One of the probable

reasons for such difference in the reliability figures of Mathematics may be that Harper et al (1967) calculated reliability for Geometry only, whereas Gayen et al (1961) and Misra (1968 a) calculated reliability for the paper containing questions on Arithmetic, Algebra and Geometry It is possible that Arithmetic and Algebra tests may be less reliable than a Geometry test Thus, the inclusion of Algebra and Arithmetic in the question papers studied by Gayen et al (1961) and Misra (1968 a) might have resulted in the low reliabilities obtained by them However, this point needs further exploration

(*b*) *Reader reliability*

As defined earlier, the correlation between the marks of two readers (examiners) is known as the reader reliability of an examination

Taylor, Thanga and Misra (1966 b) found the reader reliability of English I of B A Part I (of three hours' duration) of the Gauhati University by getting a random selection of 100 scripts (cyclostyled copies) independently marked by nineteen examiners The average correlation among the examiners was found to be 77 and the coefficient of concordance (W) was around 75 These scripts were previously marked in the University examination The coefficient of correlation between the University marks and the average of the nineteen examiners' marks was found to be 67 The low correlation of the University marks with the average of the nineteen examiners' marks seems to suggest that marking in real examinations is more erroneous than in experimental situations

Harper (1966 b & 1967 a) found reader reliability of a History paper of class X of a Secondary Board of Education by getting facsimile copies (by Xerox process) of ten scripts fairly representative of all levels of ability, marked by ninety experienced examiners of the same Board The average reader reliability was of the order of 83

Harper et al (1967) found the following figures of reader reliability for different subjects of class X examination of a Board of Secondary Education

Subject	Range of reliability figures	Median of reliability
Hindi	61—94	79
History	49—93	77
Biology	67—91	80
Mathematics	81—99	96

TABLE SHOWING RELIABILITY FIGURES OBTAINED IN INDIAN STUDIES—(*Contd*)

Content Reliability	Geography	Biology	Language	Social Studies	Gen Science	Science	Classical language	Arithmatic	Algebra & Geom	Civics	Phy I	Phy II	Phy I+II	Chem I	Chem II	Chem I+II	Aggregate Total
Gayen et al											80 (X)	80 (X)	83 (X)	75 (X)	75 (X)	80 (X)	
Harper et al		41—64															
Misra	49 (X)																
Misra																	
Reader Reliability Taylor et al																	
Harper																	
Harper et al		80 (X)															
Total Reliability Gayen et al	35 (X)																
S S C. Board Maharastra	40- 54 (X)		49 51 (X)	46- 52 (X)	54- 62 (X)	57 60 (X)	68 69 (X)	61 70 (X)	71 78 (X)	33 53 (X)							76- 80 (X)
Harper et al		50 (X)															
Misra																	

It may be observed here that reader reliability figures reported in Indian studies compare well with such figures reported in foreign countries For example as reported earlier reader reliability was found to be 31 by French (1961) and 85 by Lehtovaara (1966) This shows that Indian examiners are not less efficient than their counterparts in most advanced countries

(c) *Total reliability*

As defined earlier the total reliability is the correlation between the marks given by two parallel examiners on two parallel essay tests The present researcher is not aware of any study in India which strictly satisfies the conditions for the determination of the total reliability of the essay test However, there are studies comparing the marks obtained by a group of students on internal and external examinations both of which are supposed to be measuring the same content These results may be roughly taken as an estimate of the total reliability of the essay type tests It may be noted that the studies reported under this head have not used the term 'total reliability' They have been grouped by the present researcher under this head as to him they seemed to satisfy the conditions of the total reliability

The S S C Board Maharashtra (1960) did a similar study by correlating marks of school test examination with class X final examination marks For the purpose of sampling the examination area was divided in two regions A fairly random sample was drawn from each region The results obtained are reported below

Subject	I Sample		II Sample	
	N	Reliability	N	Reliability
Language—Higher Level	392	49	566	51
Language—Lower Level	379	57	562	47
Social Studies	255	46	408	52
General Science	374	64	500	62
Elementary Mathematics	155	79	310	69
English	327	69	432	69
Science	237	57	263	60
Classical Language	224	69	332	69
History	96	52	100	53
Geography	117	40	240	64
Indian Administration and Civics	95	33	136	53
Arithmetic	88	61	27	70
Algebra & Geometry	145	71	104	78
Aggregate Total	383	80	567	76

Gayen et al (1962) found the following correlations between the school test examination marks and Secondary Board examination marks of the same group of students in different school subjects

Subject	N	Reliability
English I	507	61
English II	636	63
English III	634	44
English Total	494	73
Bengali I	548	39
Bengali II	588	51
Bengali Total	523	55
Mathematics	483	60
History	636	43
Geography	640	35

Harper (1966 a) has statistically shown that the total reliability of an essay-type test is equal to the product of its reader and content reliabilities Accordingly Harper et al (1967) multiplied the figures of the content reliability calculated by analysis of variance method with the figure of the reader reliability obtained for the same set of scripts The median total reliabilities for various subjects were found as noted below

Subject	Total Reliability
Hindi	58
History	35
Biology	50
Mathematics	72

Following Harper s approach, that is, by multiplying the figures of the reader and content reliabilities, Misra (1968 b) found the total reliability of English I of P U Examination (of three-hour duration) of the Gauhati University to be 64

Empirical studies done by the S S C Board, Maharashtra (1960), Gayen et al (1961), and Datta (1967), and theoretical analysis by Vernon (1940) and Harper (1965) indicate that the total of marks of all the papers of an examination is more reliable than the marks on a single paper

Regarding the minimum reliability coefficient expected from a good test, Guilford (1956, p 146) observes

Following the leadership of T L Kelley there has been a general tradition that to be sufficiently reliable for discriminating between individuals a test should have a reliability coefficient of at least 94 Some have been more liberal in this regard allowing a *minimum* of 90 while others have been more demanding with a requirement of a *minimum* of 96 While these qualifications mentioned regarding reliability and validity need to be made the fact remains that in practice we expect reliability coefficients to be in the upper brackets of *r* values usually 80 to 98

It has already been explained that for practical decision one should consider the total reliability of essay tests In view of the remarks made by Guilford it is evident that *none* of the total reliability figures reported in Indian studies reach even the *minimum* limit of reliability usually acceptable in practice

4 VALIDITY

As defined earlier a test is valid if it measures well what it is supposed to measure Statistically validity is the correlation of a test with a criterion Studies on the validity of marks are generally confined to the predictive validity of marks for achievement in some other examination

Taylor (1962 b) correlated the marks of 120 college students in Chemistry Physics Chemistry Practical and Physics Practical The coefficient of correlation (*r*) and the standard error of the coefficients (*s*) were found as reported below

Physics Theory vs Physics Practical	$r = 0.14$	$s = 0.09$
Chemistry Practical vs Physics Practical	$= 0.14$	$= 0.10$
Physics Theory vs Chemistry Theory	$= 0.30$	$= 0.09$

The last correlation was affected by the presence of two very poor students If they were excluded the value of *r* falls to 20 which is barely significant He concluded that if marks are valid measures of ability there is no apparent connection between practical and theoretical abilities in Physics none between the two practical abilities and none between the two theoretical abilities As such an assumption is unacceptable we must conclude that marks are not valid measures at all

Patel (1962) in an analysis of B Ed examination marks found the correlation between the practical annual examination and the year's work in practice teaching to be 62 The correlation between total marks in Part I and Part II examination was found to be 63 Assuming that the same ability underlies the various papers of the examination his study suggests moderate validity of examination marks

Kamat and Deshmukh (1961) inter-correlated the aggregate marks of a group of students obtained in different examinations. They found the following correlations:

	S.S C.	Inter Arts	Inter Science	B.A.	B.Sc.
S.S C.		.60 N=296	.61 N=1143	.49 N=192	.21 N=370
Inter Arts				.63 N=182	
Inter Science					.42 N=347

These correlations seem to suggest a moderate validity of examination marks.

Misra G.S. (1964 a) in a similar study found the following *r*'s.

	H.S.	Inter-mediate	I.A.	I.Sc.	I.Sc.Ag.
Intermediate	.71 N=461				
I A.	.73 N=125				
I.Sc	.84 N=125				
I Sc Ag.	.72 N=125				
Graduation	.83 N=461	.87 N=461			
B A.	.69 N=125		.69 N=125		
B Sc.	.85 N=125			.87 N=125	
B Sc. Ag	.74 N=125				.76 N=125

He concluded that the results of High School and Intermediate examinations have a high predictive value for achievement in University studies.

Taylor, Thuanga and Misra (1966 c) correlated the marks of 347 students in Matriculation and P.U. examinations. They estimated the correlations between the two examinations making allowance for the students who, because of failing Matriculation, did not appear for the P.U. In other words, their value of *r* gives an estimate of correlation between the two examinations if all the students who appeared for the

Matriculation examinations were allowed for the P U examination The value of *r* was found to be 72 They also estimated on the basis of the obtained correlation coefficient, that out of 15 470 students who had failed in the Matriculation examination 1,688 would have passed if given a chance to appear for the P U examination

Lele et al (1963) found the *r* between S S C (English) and P Sc (English) marks in M S University Baroda to be 77

Harper (1963 b) explored the problem of whether the essay or objective tests are better predictors of performance on an essay type examination He had an objective examination of knowledge of English poetry as the final "Test Examination" in several Intermediate Colleges The results were correlated with the U P Secondary Board s examination in the same paper The validity coefficient ranged from — 25 to 82 (The first coefficient was not significantly different from zero) In other words, the same objective examination had different validities for predicting the same external essay examination score in different institutions The difference between the validity coefficients may have been due to inherent differences between the institutions or they may have resulted from the extreme unreliability of the essay type external examination

Harper (1963 b) in another study administered two objective tests of the knowledge of English usage to 72 Intermediate students The criteria were two terminal examinations of the essay type The correlation between the two essay tests each of 2 30-hour duration was 74 The correlation between the two objective tests each of 45 minute duration was 75 Corrected for the length of the tests reliability of the objective tests was 91, which is significantly higher than the reliability of the terminal examination of the essay type The validity coefficient of the objective tests corrected for attenuation, for predicting performance on English essay examination was above + 90 Harper concluded that "the objective test is a better measure per unit time in predicting actual essay writing marks, than an essay examination of the same length' Gayen et al (1961, 1962 and 1966) have also reported validity of some papers of class X examination of the Board of Secondary Education, West Bengal To the present researcher their validity figures seemed to fit the conditions of reliability and have, therefore been discussed under the head 'Reliability'

The above studies explored validities of marks for predicting achievement in other examinations However a more important function of examination marks is to predict performance in actual life Misra (1964 b) explored this aspect of examinations He investigated the validity of Teacher Training Courses for predicting performance in teaching on a sample size of 118 male teachers of Boys' Intermediate

Colleges of Allahabad City. His study suggested that the training of teachers had not improved their teaching efficiency significantly.

It may be observed that an examination is supposed to measure the development of various abilities and skills which education aims to bring about in students. How far the present examination system measures the development of such abilities is an area almost completely unexplored in this country.

5. *DISCUSSION*

Before concluding this part of the review it seems desirable to make a note on the practical utility of "validity" and "reliability".

Validity indicates the degree to which a test is capable of achieving certain objectives. There are three types of validity generally recognised, i.e. content, criterion-related and construct. As already stated earlier Indian studies are mostly limited to predictive validity.

Validity is the most important quality of a test. However, validation of a test requires a generally accepted criterion which seldom exists So in practice a test is considered to be valid, if the questions are relevant to the curriculum and the test is reliable.

As already defined, reliability refers to the consistency of measurement. The consistency of measurement may be considered from two aspects: relative consistency and absolute consistency. (For details see Thorndike, 1949).

Relative consistency: The coefficient of reliability, which is the correlation between two sets of scores, provides an index of relative consistency. This, being a correlational concept, suffers from the inherent weaknesses of the correlation coefficient. The correlation coefficient assumes that both the measures are normally distributed, an assumption which is often not satisfied in the case of many educational tests. The farther the deviation is from normality, the more distorted is the value of r. Besides, the coefficient of reliability is influenced by various factors other than the intrinsic agreement between the two measures, e g , the length of the test, consistency of the test items, the method by which the reliability was determined, testing conditions, and the variability in the ability level of the group on which the test was administered. Thus one has always to be on one's guard before drawing any inference from the coefficient of reliability.

Absolute consistency: The standard error of measurement gives an index of absolute consistency. It is relatively independent of the group to which the test was given. (See also Harper, 1959 and Misra 1968 b) Garret (1965, p. 351) holds that the standard error of measurement is "a better way of expressing the reliability of a test than the

reliability coefficient as it takes into account the variability within the group as well as the self correlation of the test

However both the reliability coefficient and the standard error of measurement have limited value for unscaled marks For unscaled marks it is quite possible for the reliability coefficient to approach 1 00 and the standard error of measurement to approach 0 and yet on one test all the students may pass and on the other every one may fail (see also Buros 1962)

Though nothing can be suggested as a rule of thumb to assess the efficiency of an essay test as a convenient guide it may be suggested that when two essay tests of the same length are given to the same group the efficiency of the test should be determined on the basis of reliability coefficients However when the tests are of different lengths or are given to different groups it may be better to consider the standard errors of measurement In each case the comparison is strictly accurate only when both sets of marks have the same mean and standard deviation

III Suggestions for Improving Efficiency of Examinations

Various ways for improving examinations have been suggested by experts and studies on Indian examinations

1 IMPROVEMENT IN CURRICULUM

Many experts (University Education Commission 1950 Secondary Education Commission 1952 University Grants Commission 1961 Gayen et al 1961 Examination Committee 1962 Hill 1964 a 1964 b Standards Committee 1965 Hill 1965 1966 b 1966 c University Education Commission 1966 Hill 1967 a Kaul Adalat 1970 Kadiyal 1970 Patel 1970 Ram Krishan 1970) hold that objectives of education and purposes of examinations are not well co ordinated in India Thus to bring about effective improvement in examinations curriculum must be improved Some (Secondary Education Commission 1952 Gayen et al 1961 Secondary Education Committee Assam 1965 Hill 1967 b 1967 c 1967 f 1967 g) hold that Secondary Education curricula should be so improved as to cater to the needs of various types of students (a) those who would enter life after passing examination (b) those who would enter technical education and (c) those who would enter higher education Hill (1966 c) holds that it appears that the fact that English is a second language for a great majority of students is overlooked in framing objectives for the teaching of English Hill (1966b 1966c 1967g 1967f and 1967h) recommends that examinations should measure ability to reason think interpret and apply knowledge to new

situations rather than the ability to reproduce memorised facts This seems to suggest that the curriculum should be clear on what abilities are to be developed by teaching a particular course content

2 *IMPROVEMENT IN QUESTIONS*

(*a*) *Specificity in questions*

Cieslak et al (1959) recommended that the essay question must carefully define the problem so that students need not guess what the examiner wants, the basis of grading should be clear to the examiners and the students

Ray (1970) recommends that questions should be improved so that they measure the educational objectives set by the curriculum To be specific, a question should have no room for different interpretations by paper setter, examinees, and examiners as to what should constitute the answer to the question

Specificity in questions has been explored by some studies in India Gayen et al (1961), Lele et al (1962 a, 1962 b), Hill (1967 f), and Misra (1968 b) found that some of the questions asked in our examinations are vague Such questions encourage guessing on the part of the students and lower the reliability of the test Hill (1967 f) states that in the traditional question papers there is usually no clue, except the number of marks allotted to the question, to indicate the length and scope of the expected answer Such questions give no stimulus to precision in learning

Lele et al (1962 a, 1962 b) found that there are questions which are considered to be simple by students and difficult by teachers They concluded that it is probably this differential difficulty of questions as seen by teachers and students that makes essay examinations less reliable

Some experts have studied the problem of the language of question papers Gayen et al (1961) state that ' the present system of teaching a subject in one language, but setting of questions in another creates confusion in the candidates, so the media of instruction, of paper setting and of examination (answering) must be one and the same' The Secondary Education Committee, Assam (1965) issued a questionnaire inviting teachers' opinions on this point i e , whether there was any substance in the complaint that the practice of setting questions in English while the medium of instruction was the regional language was responsible for the large percentage of failure in the examinations Some of the Headmasters replied in the affirmative Others did not agree with this view One Headmaster answered "A Matriculate or a High School candidate is expected to have at least this much knowledge of English as to understand the meaning of a question set in English

language " It may be observed that however important the ability to comprehend the English language may be, this ability should not be allowed to confuse the evaluation of other abilities There is absolutely no justification to fail a student in Geography, because he is weak in English

(*b*) *Increase in the number of questions*

The problem of having a limited number of questions in the question paper has been studied by some experts (Harper, 1960, 1962 1963 b, 1963 d, Hill 1965, 1967 f, 1967 g 1967 h, Misra 1968 b) They hold that because of limited number of questions in the traditional-type question paper, a high chance factor enters in the evaluation of achievement They recommend the inclusion of more compulsory questions in the question paper, to force a wider coverage of course content Harper (1960), Hill (1967 f, 1967 g, 1967 h) Misra (1968 b) and Trivedi (1970) recommend the use of short answer type questions in the question paper, as they feel that there is no evidence to show that the essential characteristics of essay-type tests cannot be measured by questions requiring answers in one or two sentences, or a page at the most

(*c*) *Reduction in the number of optional questions*

One problem which has often engaged the attention of teachers in India and abroad is whether optional questions should be allowed in examinations or not

Ebel (1965) recommends that giving choice among questions should be avoided unless special circumstances make such optionals necessary Other foreign experts (Stalnaker, 1951, Remmers and Gage 1955 Noll, 1957, Odell, 1958, Wood 1960, Thorndike and Hagen 1961, Lindavelle, 1961) hold that if different students answer different questions the basis of comparability of marks is weakened and this would lower reliability of marks

There are certain studies in India which have explored the problem of optional questions Taylor (1962 d and 1963 b) holds that optional questions should not generally be allowed in question papers Sharma (1963) suggests the provision of choice within a question in such a way that the alternatives do not provide a different type of activity for the students D'angel (1964) advocates abolition of optional questions on the following grounds That the candidate who is expected to answer questions is also required to select them in a hurry That the proccess of selection results in loss of time That under the strain of examination a student may not make wise selection and thus lower down

and Training to show how to write good objective-type questions Despite these recommendations, so far there has been little progress towards the introduction of objective-type tests in our examinations At the University level they have not been given even a fair trail in India

3 *IMPROVEMENT IN THE ACCURACY OF MARKING*

Taylor (1963 a) George (1964) and Harper et al (1967) found that examiners have a tendency to boost up borderline cases to the next higher division, which Taylor (1963 a) calls "J-effect" They hold that the J effect introduces further errors in marking and, therefore, suggest the need of impressing upon the examiners not to boost up borderline marks

Taylor, Tluanga and Misra explored the factors that lower marking reliability of examiners Taylor and Tluanga (1965) found that there is a positive persistence effect, that is the following mark is almost always attracted towards the preceding marks in marking The occurance of negative persistence, that is, the following mark being repelled by the preceding marks, is not disproved, but if it occurs it must be very rare Taylor, Tluanga and Misra (1966 a) found that almost half of examiners are subject to large variations of standards in the course of their marking The study showed the existence of a diurnal fluctuation in the standards of marking Taylor, Tluanga and Misra (1966 b), found that examiners' markings are much more accurate early in the process after some twenty five scripts have been marked It becomes most inaccurate about two thirds of the way through Strict examiners were found to be more accurate in marking Tendency to mark systematically low or higher was found to be a constant characteristic of an examiner The tendency to mark on a compressed or wider scale was found to be another constant characteristic To take care of the above factors in marking, Misra (1968 b) recommended the provision of model answers to examiners elaborate instructions for marking, and use of question wise marking rather than script-wise marking Harper (1963 c) showed statistically that question wise marking will reduce errors in marks Gayen et al (1961) recommended the provision of model answers and elaborate instructions for marking, and giving more time to examiners or fewer numbers of scripts to them The S S C Board, Maharashtra (1963 b) recommended training of examiners, and increase in time allowed for marking and in the number of scripts re valued by the Head Examiners

In certain universities two examiners are appointed to mark the same scripts independently When the marks given by two examiners differ widely a third examiner is appointed, and the average of the two marks which are closest are given to the student Taylor and Tluanga

(1964), in a statistical study, found that the average of the two marks which are closest may be more erroneous than the average of the two original marks that differed widely They hold that the improvement in the accuracy of marking when there are three examiners instead of two is only nominal They recommend that if a script is referred to a third examiner, the average of the three marks should be given to the student An empirical study done by Taylor, Tluanga and Misra (1966 b) supports the above finding

Iyer (1970), Shah (1970), and Loknathan (1970) recommend use of grading in place of marking

4 *SCALING OF MARKS*

Some Indian studies have explored variability in marking in different papers of an examination, and suggested methods of scaling to make achievements in different papers comparable Mahalanobis and Chakravarty (1934) found considerable difference in the distribution of marks in various papers of the School Leaving Certificate examination of 1919 of the U P Board They recommended the use of scaling marks of different papers to the same mean and dispersion by the standard score method Hossain (1940) also emphasised the need of scaling suggesting the use of a ranking procedure Bose and Choudhury (1955) recommended the use of percentile scaling Gayen et al (1951) found that even the paper-setter does not know the difficulty levels of questions Their study suggests the need of item scaling before the marks on different items are added Gayen et al (1961) and Dandekar (1968) recommend the use of equivalent scores before combining marks of different papers They hold that if scaling cannot be done for all the students this should be done at least for those whose names are to be included in the merit list Taylor (1962 e) recommends the use of scaling to equate the marks awarded by examiners differing in standards of marking Taylor (1963 a) found the use of median scaling easily applicable to large scale examinations Taylor and Tluanga (1963 b) prepared a scaling table for scaling raw marks Harper (1963 a) recommended the use of non linear scaling by fixing points at various pass divisions and fail levels Tluanga et al (1964) found that the method of scaling as proposed by Taylor (1963 a) legitimately increases pass percentage in examination because it reduces failures due to measurement error Harper (1963 b) showed that in the U P Intermediate Examination for a particular year, 64 marks of Biology are equivalent to 82 marks in Mathematics, whereas 28 marks of Biology are equivalent to 17 marks in Mathematics He recommended the use of scaling marks to the same mean and standard deviation before marks of

the different papers are compared or added. Hill (1965) also recommended the use of scaling before marks of different papers are compared. In an empirical study, Taylor, Tluanga and Misra (1966 b) found that Taylor's (1963 a) method of scaling reduces error of marking. It may be observed that despite so many recommendations put forth by various experts on examination for scaling marks, no public examination body in India except the Gauhati University employs any kind of scaling in its examinations.

5. *GRACE MARKS*

In almost every Indian examination some grace marks are given to borderline cases so as to enable them to pass the examination. The award of grace marks is often arbitrary. On statistical analysis Taylor and Tluanga (1963 a) found that such a procedure is unscientific. They recommended the use of "passing probability" instead of awarding grace marks. Those who are above a specific passing probability should be allowed to pass. Assuming standard error of marking to be 5% and 7%, they prepared charts giving calculated values of passing probability for a given raw score. Bora (1963) observed that the use of the passing probability charts should only be made after the standard error of measurement in each paper has been determined by further research.

6 *CHANGE IN THE BASIS OF AWARDING "PASSES"*

Gayen et al. (1961) hold that as the difficulty level of a question cannot be determined before students' performance on the question is known, the maximum marks allotted to each question should be given only as tentative in the question paper. The final fixation of maximum marks should be determined after the tabulation of students' performance, in a meeting of paper-setters and examiners. If this procedure is adopted, the chance factor in achievement due to selection of items will be considerably reduced. They also recommend that the minimum marks for getting a division should also be determined each year on the basis of students' actual performance that year.

Hill (1964 b and 1967 c) recommended that for the students who would enter life after passing the secondary school examination, pass should be the rule and failure an exception. He wanted that we should re-work our standard of pass, and a pass in the secondary examination should not mean a ticket for admission to university.

7. *ABOLITION OF SUPPLEMENTARY EXAMINATIONS*

Taylor (1962 f), in a statistical study, found that the supplementary

examinations enable a large number of undeserving students to pass, whereas the number of deserving students passing through such examination is only nominal He recommended the abolition of supplementary examinations

8 *INTERNAL ASSESSMENT*

To improve the reliability of traditional type tests some (Secondary Education Commission, 1952, Ministry of Education (Ed) 1956, Gayen et al 1961, Fletcher 1961, Vice Chancellors' Conference, 1962; Corey, 1964, Secondary Education Committee, Assam 1965, Mahajani, 1970, Mukerji 1970, Singh, 1970, Rao, 1970, Sarwate, 1970, Chaturvedi and Chaturvedi, 1970, Chatterjee, 1970, Sivayya 1970, Chavan, 1970, Gayen, 1970) suggest a system of giving some weight to the internal assessment of the students There are several studies which have tried to explore the efficiency of internal assessment

Pandharipande (1962) found that internal assessment is quite dependable Jain (1962), Raina (1963 and 1967), and Mallick (1968) found that there is a tendency in the internal assessment to give liberal marks They also found that in the internal assessment low marks are seldom given This makes the standard deviation of the internal assessment smaller than that of the external assessment While Raina (1967) recommended that no weight should be given to internal assessment, Mallick (1968) recommended the use of some statistical methods before internal and external assessment marks are combined

Some studies (S S C Board, Maharashtra, 1960, Gayen et al, 1962, Lele et al 1962 a, and Patel, 1962) found that internal and external assessments vary significantly in a majority of the papers These studies suggest the need for scaling internal and external assessment marks to the same mean and standard deviation before the two are combined

Misra (1969) found that internal assessment is more consistent from year to year than the external assessment in awarding marks to students

Some studies have tried to analyse the effect of combining the internal assessment with the external assessment The findings of these studies are inconclusive and sometimes controversial Taylor (1962 a) recommended that the two assessments should not be combined before verifying their accuracies Without this, it is probable that the total of the two assessments may be more erroneous than a single assessment He recommended that only the assessment which is more accurate should be considered for giving credit in university examinations Lele et al (1962 a) found that combination of internal assessment marks with the external assessment marks does not significantly improve the

prediction of further performance of students. Gayen et al. (1962) found that the total of test examination and external examination marks is more reliable than the external examination marks taken separately.

Hill (1965, 1966 a and 1967 h) recommends making internal assessment systematic, and suggests that a student's performance in internal assessment should also be reported along with his achievement in external examination.

9. *SEMESTER SYSTEM*

University Grants Commission (Ed.) (1969); Elehance (1970); Golay (1970); Desai (1970); Godbole (1970); Shapeti (1970) recommend the introduction of the semester system.

10. *VIVA VOCE*

Ray (1970) recommends supplementary essay test with *viva voce* test.

IV Conclusions Suggested by Indian Studies

As is evident from the foregoing discussion, various problems of examinations have been covered in Indian studies. However, there is a pressing need for further research. To quote Dave (1966, p. 36):

> Although several major dimensions of the field of educational measurement and testing have been tackled.....these dimensions have not yet been exhaustively studied. Much more research work is necessary in practically every aspect of the field in order to create sufficient knowledge, develop new techniques and produce effective material which may ultimately result in the qualitative improvement of educational testing in India.

None the less, the following broad conclusions may be drawn from Indian studies:

1. There is high wastage and stagnation in Indian education.
2. The standards, in terms of passes, of supposedly equivalent examinations differ considerably from one examining body to another in India.
3. The standard of marking varies for different papers of an examination, for different examiners of the same paper, and for the same examiner on different occasions.
4. None of the reliability figures reported in Indian studies reach even the minimum acceptable for grading individuals. Thus, examinations are not fulfiling their avowed purpose to "assess merit" accurately
5. The validity of essay-type tests is very low for predicting job

performance and moderate for predicting achievement in other examinations

6 There is no extensive study comparing better methods of testing with the traditional essay type tests What little research has been done in this line indicates the superiority of the objective tests over traditional essay type tests as regards both reliability and validity

7 No conclusive evidence is accumulated in favour of or against giving credit to internal assessment of the students in the final examinations

Reform Consciousness in India

The studies done in India and elsewhere have had their impact on the thinking of Indian educationists Several Government of India reports and resolutions have voiced the need for examination reform Among the early reports the most conspicuous is the University Education Commission report of 1902 However during the British regime in India Indian education was receiving inspiration mainly from the British system of education where essay type tests were still dominant Therefore nothing concrete was done towards the reform of traditional tests

The University Education Commission (1950 p 327) strongly criticised traditional essay type tests They observed

> For nearly half a century examinations as they have been functioning have been recognised as one of the worst features of Indian education in India We only note that while the magnitude of the problem has been growing at an alarming rate nothing constructive in the way of reform has happened

They further observed

> We are convinced that if we are to suggest one single reform in University education it should be that of the examinations The essay type of examination is invalid inadequate subjective and therefore not reliable

The Secondary Education Commission (1952) recommended the use of a thorough cumulative record on the basis of improved techniques of evaluation They observed (p 158) In order to maintain cumulative records properly the teacher will have to use a number of tests of different kinds—intelligence tests aptitude tests and others

The University Grants Commission's report on *Evaluation in Higher Education* (1961, p 21) observed Scholastic aptitude tests have been developed to a high degree in many other countries Such tests could form a valuable supplement to present achievement examinations

Examination Committee of the University Grants Commission

(1962, p 34) observed there are other aims the attainment of which are better evaluated by multiple choice questions, open book examinations short answer examinations *viva voce* tests, etc"

The Standards Committee of the University Grants Commission (1965 p 79) observed The use of objective tests will not only make internal assessment more reliable but will also ensure speed, economy and efficiency

The Education Commission (1966, p 291) recommended the introduction of objectivity in our examinations They observed 'Another important point of emphasis would be the reorientation of university teachers to adopt new and improved techniques of evaluation'

It is apparent from these recommendations that there is a great awareness in India of the need for examination reform Only the desire for change has got to be felt thoroughly enough at the lower levels so that the change becomes possible

Agencies Active in Examination Reform

As a result of the foregoing many agencies and educational bodies in India have become active in the field of examination reform Among the most important of these are the Department of Curriculum and Evaluation NCERT, National Institute of Education (NIE), Evaluation Units of Secondary Boards, Gujarat Research Society, Bureau of Educational Research, Ewing Christian College, Allahabad, Bureau of Psychology, U P the Kerala Baroda Mysore, Gauhati Universities, the Indian Statistical Institute (ISI) Calcutta and the Regional Colleges of Education Of agencies working on public examinations the Gauhati University among the universities and the Central Examination Unit are in the vanguard (Dave, 1966, p 10)

Besides several important studies were made under the grant in aid schemes of the National Council of Educational Research and Training and the University Grants Commission Among the studies done under the grants in aid schemes those done by Gayen of Kharagpur Harper of Allahabad, and Pillai of Kerala are outstanding

A brief report on the progress of examination reform in India is contained in DEPSE (1967 b) and (1967 d)

Publications Pertinent to Examination Reform

Publications in the field of educational research have appeared in great number in recent years especially from the work of the above mentioned groups A few very important ones dealing specifically with problems of examination reform are mentioned below

Publications of NCERT

The Concept of Evaluation in Education
Improving Examinations
Optional Questions in Tests and Examinations
Dimensions for Setting Questions in Board's Examination
Evaluation in General Science
Evaluation in Physics
Evaluation in Geography
Reports on all India Workshops on the Development of Test Material
Booklets on specimen test items in various school subjects
Besides, *Examination Abstract* Nos 1 to 6 published by the Department of Evaluation and Curriculum give valuable information about research studies abroad and in India

Publications of the Gauhati University

Operation Passmark
Grace Marks
Scaling Tables
The Influence of Scaling on Examination Results
The Problem of the Third Examiner
The Persistence Effect in Marking
The Inconstant Examiner
A Study in Multiple Marking
Examinations As Predictors
An Investigation into Admission Criteria
A Study of Essay Type Test Reliability With Some Suggestions for Its Improvement
A Study of the Effect of Randomisation and Scaling on The Errors in Examination Marks
Difficulty of Essay Questions
Discrimination Values of Essay Type Questions
Effect of Examiner Variability on Difficulty Indices of Essay Items

Other publications

Specimen test items prepared by the Mysore Maharashtra and other Secondary Boards of Education and the Punjab University
Reports on evaluation workshops published by the Mysore Maharashtra Madras Bihar and Assam Secondary Boards of Education, and the Meerut University
A Handbook of Selection Use and Interpretation of Psychological Tests by A Edwin Harper Jr, published by U S Educational Foundation in India 12 Hailey Road New Delhi

Examination Reforms in Maharaja Sayajirao University of Baroda, published by the M. S. University, Baroda.

Considerable research on examinations has been done, yet much more is needed to provide essential knowledge about educational measurement. Our knowledge of the human mind is far from satisfactory. What knowledge, abilities and skills are of most worth and how to measure them best are the questions which have not been answered conclusively so far.

As regards examination research in India, it is still in its infancy stage. Probably the Examination Research Unit of the Gauhati University is now the only agency solely devoted to research on examinations. There is no periodical exclusively devoted to educational measurement. It is hoped that with the advance of time examination research will gather momentum in this country.

Besides, examination research in India suffers from various handicaps. Many of our universities do not train their students in research methodology. The result is that sometimes research studies have poor experimental design, sometimes assumptions not warranted by the data are made, sometimes some essential information is not reported which makes it difficult to draw an independent conclusion, and sometimes wrong inferences are drawn. We do not yet have even an agreed pattern of research reporting. Much research work done has not been published. A huge number of dissertations submitted for Master's or Doctor's degree in Psychology and Education is on educational measurement. A great majority of them are on test standardisation, most of which have no further utility than the award of a degree.

There are certain problems which can be handled by some central agency. Standardisation and development of national norms of intelligence and achievement tests is overdue now. The determination of objectives of university education is observable behaviour, construction of admission tests to various courses, and exploration of validities of various courses are the areas where a central agency could be of immense help. It is unfortunate that Indian universities take approximately three months' time for the publication of results. A central agency could help universities to speed up the tabulation of results by modern methods.

Despite the need for more research and improvement in the standards of research, it may be stated that enough research on examinations has been done to lead us in a definite direction Poor reliability and validity of essay-type tests is now established beyond any reasonable doubt. There is a need to try methods supposed to be better ones, at least on an experimental basis. When the results seem encouraging we should adopt these methods in our public examinations

Research by itself cannot reform examinations Research findings have to be disseminated to those who have the power to reform examinations Researchers use a technical language, and rightly so because they address their co-workers, but such writings make little sense to people at large We need researchers who may write reports in laymen s language We need other talented people, who, though not actually engaged in research, may interpret the research findings to the general public through popular articles in newspapers and magazines

Most of our teachers do not know what objective based teaching, learning and evaluation means We need to train them in the modern concept of evaluation The need of improved evaluation is there at every level of education but it is dire at the university level because it is the university students who are the future leaders of our country, and to the extent that we are able to measure their competence, to the same extent we shall be able to get competent people to lead and govern us

6

Data and Experimental Design

THE GENERAL purposes of the two experiments have already been described in Chapters III and IV It is hoped that the reader has already read those, to get an overview of the projects It is now time to give the more technical details

In this book the descriptions of the "ninety marking ten' experiment have preceded the "four thousand re examined" project This is because the former is perhaps more striking and easy to understand However, it is easier to describe the data gathered for the latter study first, as it was the more extensive The few ways in which the data of the "ninety marking ten" project differed will then be given

A Four Thousand Re examined

1 Examinations selected

The original purpose of the study was to answer two related, though somewhat different, questions "How reliably can an examiner re mark answer books previously marked by himself? ' and ' How reliably can an examiner re mark answer books previously marked by another examiner? ' (As we shall see, new problems arose out of the data itself, so that ultimately a good deal more than just these two questions was answered)

There are two possible approaches to this sort of study The first is to set up a special examination experiment, asking students to answer

a question paper, and having each answer book marked by two examiners The second is to use the results of an actual examination administered under the usual conditions

One big advantage of the first approach is that two *completely* independent markings are possible i e the first examiner can be instructed to make no marks in the answer books However, the type of answers that students may write in such a situation may be different from those written in a regular examination More important examiners under such circumstances may act differently from the way they usually do, and thus the results may not be believed to apply to the routine situation Nevertheless the first approach has considerable merit and, after finishing this study, the junior author went on to carry out a study along those lines In the present study, however, actual examination answer books were used

The generous cooperation of a Board of Secondary Education was obtained The Board which provided the basic data has asked to remain anonymous on the very reasonable grounds that the results of this study are probably applicable to *any* Board, and thus it would be unfortunate if any particular Board (rather than the whole examination system itself) were stigmatized because of these findings (*Those who may already know the identity of this Board are asked to respect this request for anonymity*)

The Class X Final Examination was chosen as it is one of the most widely administered and crucial general examinations in the entire educational system

Four subjects were chosen *History, Hindi, biology mathematics* These were chosen as fairly representative of the total system. There are two Arts exams, one of them social studies and one language There are two Science exams one involving considerable judgement in marking and the other a supposedly more exact science For simplicity, only one paper, generally the First Paper, was chosen to represent each subject In mathematics the Second Paper was selected as this was the geometry paper which was considered to involve slightly less mechanical marking than the First Paper Each of the four was a three-hour paper The Hindi paper carried 34 marks, the others 50 marks each

2 Examiners selected

The Secretary of the Board was asked to provide answer books from eleven experienced examiners in each subject, i e the work of a total of 44 examiners The examiners in each subject were selected from among those who had worked under the same Deputy Head examiner, and thus could be expected to be experienced in the same method and standard of marking

3 Experimental design

One thousand answer books in each subject were selected Each examiner re examined 50 answer books originally examined by himself, and 50 originally examined by another examiner The examiner was not informed as to which were originally marked by himself, and which originally by another He was informed that "some of them were originally marked by yourself, and some by another examiner", but not how many scripts were of each type

The selection and redistribution of answer books was made according to the following design

		Scripts originally marked by											Total re-examined by each
		A	B	C	D	E	F	G	H	I	J	X	
Examinar													
0	A	50	50										100
1	B	50	50										100
2	C			50	50								100
3	D			50	50								100
4	F					50						50	100
5	F						50					50	100
6	G							50				50	100
7	H								50			50	100
8	I									50		50	100
9	J										50	50	100
Number originally marked by each		100	100	100	100	50	50	50	50	50	50	300	1000

Thus each examiner re marked 50 answer books originally marked by himself and 50 originally marked by another examiner The first two examiners (A and B) were paired, each re examining 50 of his own scripts and 50 of the other's The second two examiners (C and D) were likewise paired The remaining six examiners re examined their own scripts and those of the eleventh original examiner, Examiner X Examiner X was specially chosen as one who had had considerable experience in marking (It was originally specified that Examiner X should also be one "whose marking has conformed closely to the normal pattern" Without previous knowledge of the means and standard deviations of the marks this specification was impossible to apply As it turned out, three examiners did conform fairly well to this requirement, but the mathematics Examiner X's mean was considerably above average for his subject)

4 Selection of answer books

The Board arranged to have answer books drawn from their stacks, before the 1964 examination booklets were destroyed From among these, the Research Assistants selected the requisite number to fulfill our experimental design

The plan stated that the scripts of each examiner will be so selected that they include a reasonable range of his marks' At the Board s offices the answer books are kept in large cloth bags in huge stacks in the store rooms Though specific books could be found with much searching, apparently, they could hardly be considered to be in 'apple pie order" Thus a strictly random sampling would have required an enormous amount of searching and other work, an amount which did not seem justified under the present circumstances Were this a study of the statistics of examinations (as, for example were Gayen s 1961 and 1962), then a strictly random sample would be needed But since this is a study of re examining, it is only necessary to be reasonably certain that the selection of answer books was not done in such a way as to bias the results of the re-marking experiment

In general, the Research Assistants selected the first answer books that they could find, of each of the 44 examiners Since the answer books were not stored in any particular order, there is no reason to believe that the ' first found ' are *not* essentially a random sample

Is it possible to test the representativeness of this sample? Unfortunately, the data required for such a test are just not available The Board compiles no statistics except "pass percentage' But even this is of very limited use to us, because (a) it is based on all candidates, while we limited our study to regular (disregarding private) candidates, (b) it is based on the total subject, and no separate "pass percentages" are available for the separate papers, and (c) something like 7% of the Board's candidates pass by "grace marks", which are not included in our study

With all these reservations, let us examine such pass percentage data as we have available

	History	Hindi	Biology	Mathematics
Board (all candidates)	75 0%	77 0%	75 4%	64 6%
Our study (regular candidates only)	74.5%	70 9%	88 2%	60 7%
Differences	—0 5%	—6 1%	—12 8%	—3 9%

As we have noted, the Board's pass percentage is based on the sum of two (and in Hindi three) papers, while ours is based on only one of the papers, in each subject If the marks of two correlated papers are

summed, and the pass percentage is the same for each of the papers, the pass percentage on the sum will be *higher* than the pass percentage for the individual papers This follows from the formula for the standard deviation of a sum

$$\sigma_{(x+y)} = \sqrt{\sigma_x^2 + \sigma_y^2 + 2\, r_{xy}\, \sigma_x\, \sigma_y}$$

If r_{xy}=1 00, then the standard deviation of the sum is twice that of each individual paper But if r is less than 1 00 then the standard deviation is smaller than twice the original standard deviations Since the standard deviation has thus shrunk, but the mean of the sum is always twice the original means, then the standard score distance of the pass mark from the mean is *greater* for the sum than for the original two papers—and, therefore, the fail percentage is lower and the pass percentage higher

Thus our pass percentages should be lower than those of the Board—which they are in three of the subjects The fact that the difference is not greater probably reflects the inclusion of private candidates in the Board's statistics As for biology, this may mean that our sample is rather better than the average Or it may mean that Paper I was much easier than paper 2 It is equally possible, however, that our results were biased by the individual characteristics of the particular Deputy Head Examiner under whom our examiners had worked This hypothesis is given some slight credence by the fact that the mean marks droppes more in the second marking in biology than in any other subject, and that the mean marks of the "specially experienced Examiner X" were lower than the means of the ten original examiners Again, however, the Board did not provide us with any statistics as to how those who worked under one Deputy Head Examiner differ from those who worked under another, so no more exact comparison can be made

What evidence we have supports the conclusion that there is no reason to believe that our sample is not adequately representative of the total population from which it was drawn, for the purposes of a study of this type

5. *Preparation of answer books*

It was at this point that we ran into the first of several unforeseen and very difficult problems This one was so serious that its solution added more than two months to the original time schedule of the project The problem was the "disguising" of both the identities of the candidate and the original examiner, and the marks of the original examiner

It was fairly simple to design and print a new cover, so the old cover could be removed from each booklet before sending it out This removed the primary source of identification of examiner candidate, and marks

awarded In a very few instances a diagram had been drawn on the back of the cover Every cover was of course checked for this In these few instances the diagram was cop ed as exactly as poss ble nto the back of the new cover

Both the candidate s roll number and the examiner s s gnature also appeared at the end of each booklet Usually these could easily be cut out with a razor blade as there was nothing written on the reverse

Where the number of marks awarded a question was written in the margin this also was cut out with a razor Before cutt ng we had to check the reverse side If any word on the back of the sheet were cut out in this process it was re written on the remain ng paper (We experimented with various other techniques such as blacking over with Ind a ink removing with ink remover pasting over with paper etc None of these proved to be adequate although they were used in occas onal cases where cutting would have destroyed a part of a d agram)

But for the most serious problem we found unfortunately no completely satisfactory solution The previous examiners had placed red tick marks X marks underlinings etc throughout each booklet *over the candidates writing* on those places where they cons dered the candidate s response right or wrong It was impossible to remove these marks without also removing the candidate s own writing So the only possible solution remained to try to disgu se the examiner s markings as much as possible Several techniques were employed At the start tick marks were changed into X marks by add ng a red line These X marks had a tail left over from the orig nal t ck marks—so tails were added to *all* X marks But all this proved very laborious and t me consuming so a different techn que had to be adopted Wherever there was a tick mark an X mark was added beside it—and vice versa Under linings of course could not be changed—but more underlin ngs (as well as extra crosses and tick marks) were added at random Thus the re examiner hopefully would not know which were his orie nal marks and which were the new added marks Finally the exam ners were specifically asked to ignore the old markings when re marking the answer books The directions were

> It was of course necessary to disgu se the previous marks on the papers So you will f nd many ticks and cross marks in the answer books Some of these have been marked by the examiners and some by our office You should not be guided by them As much as possible it is best that you ignore them completely
>
> You may use ink of some other colour than red to identify your new marks.

We found that each Research Assistant could disguise a maximum

when booklets were arranged in strictly numerical order for sending out, this arrangement effectively mixed up the "same" and "different" examiner booklets in an essentially random pattern

The next step was to stamp the proper code number on both the old and the new cover of each booklet It was also stamped on the first page of the booklet, as a protection against possible detachment of the cover The old cover was then removed and the fresh blank cover stapled on New covers were also added to any "B' booklets The 'disguising" of the answer books was complete

7 Despatch and receipt

In the meantime, the Secretary of the Board had provided us with addresses of the examiners, and also question papers, rules and instructions to examiners, supplementary instructions of Head Examiners, etc We had written to the 40 examiners, enclosing reply post cards, and had received assurance that they would cooperate with the study If any examiner did not reply within a week, a reminder was sent All this was done before much work was done on the answer books Arrangements had been made to obtain other batches of answer books in case any examiner refused Fortunately, this was not necessary

As soon as it was ready, each batch of answer books was despatched by railway parcel (We had considered parcel post, but found it substantially more expensive) By registered post we sent the RR Bill, question paper, the Board s instructions (which we reprinted, with the Head Examiner s comments added), and a bill form for remuneration and expenses Examiners were to be paid at the regular Board rates We also enclosed a money order form asking the examiner to fill in his correct name and address—thus saving us not only extra work, but also possible error

Any examiner not returning the answer books in time received a reminder, but no deduction was made in remuneration

As soon as the marked answer books were returned, the data were prepared for card punching Each cover was removed, and matched up with the old cover The new cover was stapled onto the old cover in such a way that both sets of marks were visible, side by side, but the original roll number and other information about the candidate was completely covered This, again, was to preserve anonymity At the same time the number of pages answered in each booklet was also counted, and recorded on the cover

8 Card punching

Arrangements had been made to punch the data on I B M cards at

the Life Insurance Corporation offices in Kanpur The 4 000 pairs of marked covers were taken there personally by the junior author the requirements explained and the final card design determined His presence was very necessary throughout the early part of the work as unforeseen problems required new adjustments For example Questions 9 and 10 in History were of the short notes variety with many alternatives For item analysis purposes we had originally wanted each alternative punched separately But it was soon d scovered that many examiners had not given information on the cover about these separate alternatives More serious it was found unexpectedly that some examiners had used quarter marks on some quest ons To keep the data for a pair of examiners within a single card we had allowed only three columns for each question carrying 10 or more marks The third column was to allow the addition of 0 for a whole number or 5 for a half Thus 8 1/2 would be punched 85 It was not unt l a large number of cards had already been punched that some quarter marks began turning up Rather than re-do the entire job for these few it was considered better just to round the quarter marks to the next higher or lower half mark while making sure that the total still added up correctly For biology for which punching had not yet been started allowance was made for the quarter marks—though at the cost of doubling the number of cards needed

Board instructions required the rounding of half marks in the totals to the next higher mark However we punched *both* the half mark and the rounded total For most calculations the rounded total was used as this was the official mark However, for the content reliability formula the actual totals not the rounded ones were needed for accuracy

9 Computations

Computations were performed on the I B M 1920 at the Indian Institute of Technology Kanpur The junior author had kindly been invited to take a course in FORTRAN programming and was thus able to handle the entire job

Again in planning this research project due to our inexper ence we entirely misjudged the time that would be requ red for the com puter It is true that the computer calculates very rapidly It is somewhat slower in feeding large quant ties of data such as we were dealing with More important because of its popularity it is about as accessible as a film star! At most we could have it for a maximum of a half hour in any given day Furthermore it is temperamental and frequent breakdowns stretched out the time st ll more It required more than

a month to complete the computer work, which we had expected to be done in only a few days.

Furthermore, this particular computer has a rather limited memory. The complete analysis of the results of each individual examiner was within this capacity. But the summary charts which form the bulk of this report all had to be tabulated by hand. This again added enormously to both the time and effort required, far beyond our original expectations. Most of this has had to be done by the senior author himself, without help, since his return from the U.S.A.

However, it is also true that the project might have taken years, rather than months, *without* the computer. The 240 coefficients of correlation alone would have taken several months, not to speak of the means, standard deviations, scatter charts, and various frequency distributions prepared for each of the 80 sets of 50 answer books.

Each set of calculations was done by hand for one or two sets and checked against computer results, to check on any systematic machine errors due to programming error. Further, the results were scanned, and any very unusual results double-checked. (Some punching errors were found—despite the fact that the cards were supposed to have been verified.) The computer was programmed to calculate up to four decimal places, rather than just dropping digits beyond the fourth place.

Appendix E lists the information printed out for each set of 50 answer books, not all of which is reported or even used in this report. Note, especially, that sums and sums of squares are available for analysis of variance which (for lack of time) has not been done.

10. Formulas used

Standard formulas for mean, standard deviation, and product-moment coefficient of correlation were used. N, and not N-1, was used in the denominator of the σ formula. In calculating t for the difference between means, the formula for correlated means was used, and correlation between the two examiners was used as an estimate of the correlation between the means.

It was discovered (after all the calculations were finished) that the t-test for difference between correlated standard deviations does not apply when N's are less than 100. The F-test, of course, does not take account of correlation. Dr. C. R. Rao, Professor and Head of the Research and Training School, Indian Statistical Institute, Calcutta kindly provided us with the following significance test for this situation.

$$r = \frac{S_{xx} - S_{yy}}{[(S_{xx} + S_{yy} + 2S_{xy})\,(S_{xx} + S_{yy} - 2S_{xy})]^{1/2}}$$

where S_{xx} is the corrected sum of squares for x variable (scores of examiner A) and S_{yy} is the corrected sum of squares for variable y (scores of examiner B) and S_{xy} is the corrected sum of products The statistic r behaves like the correlation coefficient based on a sample of size 50 (The letter, from which the above is quoted, goes on to explain that d f for our sample is 48, and the table of significance of r is used Dr Rao also kindly supplied a numerical example)

A few days later we also received the following solution from Dr A B L Srivastava, Department of Psychological Foundations Delhi In a later communication, Dr Srivastava informed us that this test is in the Third Edition of McNemar's *Psychological Statistics*

$$t = \frac{(s_1^2 - s_2^2)\sqrt{N-2}}{\sqrt{[4s_1^2 s_2^2 (1 - r_{12}^2)]}}$$

which is distributed as Student's t with N-2 d f under the null hypothesis Here s_1^2 and s_2^2 are sample variances and r_{12} is the sample correlation

Dr Srivastava goes on to state (in a later letter) "You will probably find it interesting that the solution suggested by Dr C. R Rao yields the same result as the t test suggested above Dr Rao has suggested testing significance of r, which can be done by calculating

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

"This t follows Student's t distribution with N 2 d f under the null hypothesis that population value of r is zero *or* H_0 ($\sigma_1 = \sigma_2$) Using the expression for r given in Dr Rao s letter, the second formula reduces algebraically to the first, so the two tests are identical

Dr Srivastava has shown that his formula and Rao's are algebraically related, and should produce identical results Since calculations had to be done by hand and therefore needed to be checked, the calculations were done with Rao's formula and Srivastava's was used to provide an independent check

B Ninety Marking Ten

The ways in which the "ninety marking ten project differed from the "four thousand re-examined" experiment are now to be summarized

The basic difference was this One project called for the re marking of a large number of answer books by only one examiner each The other project called for the re-marking of a small number of answer books by a large number of examiners

1. Choice of method of reproduction

There were several ways in which one hundred copies of each answer book could have been made. The simplest, of course, would have been to have had them typed and mimeographed or printed. Taylor et al. (1966) used this method, "taking every care to copy as exactly as possible. Errors in spelling, grammar, punctuation, etc., were duplicated, and words written and cancelled by the candidates were also copied." But whether all the candidate's errors can always be copied is problematical, and in any case this method becomes much more difficult where there are drawings or mathematical symbols to be reproduced. More important, the factor of handwriting is eliminated altogether— and the fact that typing or printing is so much easier to read, may influence the marks. A second method is to have someone copy the answer book by hand, onto mimeograph stencils or spirit or offset masters. This, also, is relatively inexpensive, and may produce multiple copies closer to the original. But unless one has several copiers, in fact a different one for each answer book, the factor of handwriting differences is again lost. And, in any case, copying will inevitably be neater than the original script, which is written in haste and under the pressure and anxiety of examination conditions.

The only answer to all these objections is to reproduce the original answer book photographically. At the time that this study was initiated, this was easier said than done. Photo-offset was a new process in India, therefore hard to obtain, and prohibitively expensive. Multigraph (both Multilith and Rotaprint) were available, but again the photographic process was far beyond the budget of this research. Thermofax was available for inexpensive production of masters, but this required scripts written in pencil or with a carbon-content ink, so it also was ruled out. For a while, in fact, it looked as though it might actually be cheaper to have the material reproduced abroad. But after several weeks of investigation, we found that Xerography was being introduced into India. The Xerox process is among the least expensive methods of producing facsimile multigraph or offset masters. Messrs. Gaumont Kalee Ltd. of Bombay were importing a Xerox machine capable of producing Rotaprint masters, and they already had a Rotaprint on which to make the prints. They agreed to undertake our work, as soon as the Xerox arrived.

2. Choice of examinations

The original intention was to have ten answer books re-marked by one hundred experienced examiners, in each of the four subjects used in the "four thousand re-examined" project. But both the time

limit and the budget of the project made this impossible The fact is that at the time of planning the necessary reproduction method did not even exist in India, and thus neither time nor cost could be estimated with any degree of accuracy Thus the search for a reproduction method delayed us by several months, the method itself was much slower than planned, and its cost was far beyond what we had estimated For these reasons we did not get beyond the study of the first examination selected for reproduction i e history

3. *Choice of answer books*

Ten history answer books were selected from those actually written in the Class X Final History Examination in an Indian state*, in the spring of 1964 The answer books were selected from among those who had answered the same combination of questions This gives the *best* possible situation for placing the answer books in an accurate rank order If—as is usual—we had asked examiners to compare answer books answering different combinations of questions, the judgement task would have been much more complex, and the results would probably have shown even less agreement than they did (See Harper, 1962)

The answer books were also selected to give a typical distribution of marks, so far as the original examiners' marks were concerned From the data obtained in the companion study, we knew the distribution of marks in this paper A table was prepared indicating the number of answer books desired, within each range of about half a Division The stocks in the Board office were then searched until we found answer books which contained the required combination of questions, and had marks in the desired ranges We selected the first answer books that we found that met these requirements Thus we may consider this to be, in a rough way, a stratified random sample of this Board s History answer books

4 *Preparation of answer books*

It had seemed important to select actual examination answer books, rather than to set up an artificial situation with examinees taking a new examination Thus the factor of motivation, which might be very different in the two types of situation, was kept as "realistic" as possible

However, as we saw earlier in this chapter, this introduced an unforeseen difficulty The examiners who originally marked the answer books had made red marks all over them We hoped that these red marks

*The Board which cooperated with this study has asked to remain anonymous Since the results of this study are likely to apply to any similar examination, they feel it would be unfair to allow interested or hostile persons to use these data for unfavourable publicity against any particular Board

could be eliminated by using a red filter in the reproduction process, but Gaumont-Kalee informed us that they could not do this. (For future investigators let us add this note: Photo-offset is now much more widely available in India, and we are informed that a filter can be used to eliminate the red marks if answer-books are reproduced by the photo-offset process.) Thus we were left with no choice but to try to "disguise" the original markings. This we did by the methods described in the previous chapter. A questionnaire sent to the participants in that study indicated that the disguising was reasonably successful, so we may assume the same in this study.

Of course, we probably were not completely successful. The correlation (r) between the mean marks awarded by the 90 examiners and the number of previous markings in each answer book (including the bogus ones we added) was +.41, while the correlation of the marks awarded by the original examiners with these markings was +.29. Since we are interested in these groups themselves, and not in any statistical inference, we may consider this difference significant for these data. It is logical to assume that there would be *some* relationship. The +.29 suggests that the random addition of markings may have reduced the original correlation, while the +.41 suggests some (unconscious?) influence on the 90 examiners.

Still, we do not really know the extent to which these previous markings determined the results of this study. This is the price we had to pay for having a "realistic" (rather than an artificially-produced) set of answer books, in the candidates' own hand-writing, with errors corrected, mistakes, etc. It is obvious, however, that whatever influence these disguised previous markings may have had would have been to *increase* the degree of agreement among examiners. We may, therefore, interpret the following results as *maximum* rather than average or minimum inter-examiner agreement figures.

A code number was written on each page of each answer book, before sending to Bombay. Using these code numbers, we collated the sheets sent us from Bombay into test booklets. The reproduction took more than three times as long as the company (which had had no previous experience with this process) originally estimated, making it impossible for us to finish the project on our original schedule.

5. *Experimental design*

Dr. H. J. Taylor has drawn attention to what he calls the "persistence effect" in the marking of examination papers. We were, therefore, interested in whether the serial order of marking would influence the marks given individual answer books. Ideally, the papers should have

been marked in a different order by each examiner, but the administrative complexity of ensuring this, and the clerical complexity of tabulating results ruled this out Thus we decided on using two orders only, one the reverse of the other The ten answer books were arranged in random order, then numbered 10-A, 9-B, etc through I J Two types of marks sheets were prepared—in one of them, the books were listed in the order "A, B, C, J ', in the other the books were listed in the order "I-J, 2-I, 3-H, 10-A " The first order was sent to the first 25 examiners on our list, the second order to the second 25, the first order to the third 25, and the second order to the last 25 examiners Since the names of the examiners were presumably in no particular order, this was assumed to randomize the procedure adequately Examiners were asked to mark the books in the order received

6 Despatch and receipt

At our request, the Secretary of the Board of Secondary Education had provided us with a selected list of 100 *experienced* examiners in Class X History As in the four thousand re marked experiment, each examiner was supplied with the question paper (only those questions actually answered), the Board's and Head Examiner's regular instructions, and the supplementary intructions of the deputy head examiner A typical instruction for one of the questions '4 marks for social condition, 3 marks for political condition and 3 marks for economic condition " No model answers were supplied The instructions were the identical instructions which had been used for the regular examination All the examiners had marked this paper in the regular examination Therefore we may presume (though we have no proof, of course) that their "standard of marking" for these 10 answer books was approximately the same as for the original examination

To save both time and expense, no attempt was made to contact examiners ahead of time A letter explaining the project and requesting cooperation was sent along with the answer books, by Registered Post Each examiner was offered an honorarium of Rs 10/- for marking and returning the examination books Under these circumstances a 90% return seems reasonably good, considering that examiners may have moved, been ill, or otherwise disinclined to cooperate Repeated attempts to trace them and get back the answer books were made, of course, but ultimately the marks sheets of the remaining ten could not be obtained

7 Reliability of the data

However the fact that only 90 of the 100 examiners returned their

marks sheets does raise the question: How serious is the fact that 10 did not mark the papers ?

It is fortunate that our design gave us two separate samples of examiners—those to whom the answer books were sent in the first order, and those to whom they were sent in the second order. Forty-two of the first group returned their papers, and 48 of the second. It is reasonable to assume that the 48 are representative of the 50 in the second group. If the 42 in the first group can be shown to be similar to those in the second group, it seems reasonable to assume that the 42 are an unbiased sample of the original 50 in the first group, and that the 90 examiners therefore fairly represent the original 100.

First, the significance of the difference between the means of the marks awarded by the two sets of examiners to the same ten answer books was examined. The difference (0.31 marks out of 50) divided by its standard error was .70, which certainly does not suggest any real difference.

For a more exact test of the distribution as well as the means, chi-square was applied. The total frequency with which each mark (ranging from 2 to 38) appeared was tabulated separately for each of the two samples of examiners. Small frequencies were combined. The two distributions were then compared. They yielded a chi-square of 16.58555. For 26 degrees of freedom, the probability of this chi-square appearing in two samples from the same population is between .90—.95. For chi-square to reach even the 5% level of significance, it would have had to exceed 38.885.

Both these arguments depend, of course, on the fact that the answer books were distributed in a random order. Any interaction effects between order of examining and marks would thus tend to be averaged out. Clearly, these two samples are identical. They are both very near to complete, and the missing 10% of examiners would not have affected results significantly.

7

The Concept of "Reliability" as Applied to Traditional Examinations

CONFUSION OF terminology, in any science can lead to confusion of thinking and of interpretation—and this we fear is what has sometimes happened in discussing 'reliability (See Review of the Literature Chapter V) The APA AERA NCME *Standards* (1966) have attempted to clarify the various meanings of reliability primarily as the term relates to objective tests and examinations We have not found however any clear and complete treatment of the subject as it relates to traditional essay type examinations or to the comparison of the two types For example when looked at from the standpoint of essay type examinations the reliability of an objective test is really two different reliabilities represented by numerically identical coefficients

We are indebted to Warren Findley who after reviewing an early draft of this study, first suggested (in a personal communication) this need for terminological clarification and who suggested some of the concepts needed And as the knowledgeable reader will soon find we have drawn heavily on the work of Harold Gulliksen (1950) and of H.J Taylor However we have tried to adapt these concepts to traditional Indian examinations terminology also—e g marks instead of scores , and examiner instead of reader

First let us define a basic designation *traditional examinations* Although we have also used the more usual term essay type examinations we feel that traditional may be a more felicitous term By this we mean questions requiring long answers worked out in detail in a page or more A mathematics question requiring several pages of algebraic

transformations or calculations, but no words, is hardly an "essay"; but it is not "objective" either. It is "traditional".

The other basic concept is *scale* : We will talk frequently about different subjects, and even different examiners, having different "scales". What do we mean? The reader is familiar with two different scales of "degrees" of temperature—degrees Fahrenheit and degrees Centigrade. Similarly, there can be many different scales of "marks out of 50". A particular examiner's scale is most simply defined in terms of two statistics, the mean and the standard deviation of the marks he awards to a randomly selected batch of answer books. The relationship between a candidate's marks on the scales of any two examiners marking the same batch of answer books is given by the formula :

$$\frac{M_x - X_x}{\sigma_x} = \frac{M_y - X_y}{\sigma_y} \qquad (7.1)$$

where X_x and X_y are the marks awarded the same answer book by examiners x and y, M_x and M_y are the means of the marks, and σ_x and σ_y are the standard deviations of the marks awarded by examiners x and y respectively. (Mahalanobis, 1934 takes into account other statistical characteristics also, but the ones we have noted are sufficient for most ordinary purposes.) Once the constants have been calculated for the particular examiners (or for an examiner and an arbitrary scale), this reduces to the form $X_x = X_y + b$, which can be quickly calculated for each individual candidate. But it should be noted in passing that this formula is relevant only when marks are normally distributed. Other scaling methods, more appropriate to marks in India, will be discussed in a separate chapter.

1. *"Examiner", "content", and "marks" reliability*

Now let us start with two parallel traditional-type examinations, Form A and Form B. Let us assume two examiners, or readers, for each: Examiners 1 and 2 for Form A, and Examiners 3 and 4 for Form B. In this situation, we find three distinct uses of the term "reliability".

Examiner reliability : This is the correlation between the marks given by two separate examiners to the same set of answer books. (American texts call this "reader reliability"). Too often, as has been pointed out in the Review of the literature (Chapter V), researchers have just called this "reliability". This has then led them to the erroneous assumption that, if this coefficient can be raised to .85 or .90, "essay examinations are just as reliable as objective examinations". (See Basumallik's, 1959 review for examples.) Note that in objective examinations, "examiner reliability" (or "reader reliability") is 1.00.

Actually, one must distinguish between two types of examiner reliability here One is when an examiner re-examines answer books which he *himself originally marked* Let us call this simply the "self-consistency" of the examiner The other type is when the answer books are re-marked by a different examiner This may be called "inter-examiner consistency" In various contexts in this book, these two are distinguished by the words "Same" and "Different" (or just "S" and "D"), and "Self" and "Other" If the number of scripts is large and/or the time interval long, memory is a negligible factor in examiner self-consistency Where the two forms of examiner reliability differ, then such factors as different "expectations" or "styles of marking" are probably operating

Content reliability This term was introduced by Gulliksen in 1936 (see Gulliksen, 1950) It might as easily have been called examination reliability or *test reliability, and is the only type of traditional-type* examination reliability which is strictly parallel *in concept* to the reliability of an objective examination ("In concept' is italicized because, as will be seen later, it is more accurate *in practice* to compare 'marks" and not "content" reliability of essay-type and objective examinations) This is the reliability of the question paper itself, quite apart from the reliability of the examiners Statistically, it is the reliability of the test corrected for the attenuation due to examiner unreliability

Marks reliability This is the reliability of the total examination situation, taking into account both "examiner' and "content" sources of unreliability In fact we have previously called this "total reliability" (Harper 1967, Misra 1968) But we have recently seen the term 'score reliability" used (Coffman, 1966), and this suggested that "marks reliability" would be a more descriptive term for use in India

Note that the same coefficient represents both the content reliability and marks reliability of objective examinations But for traditional examinations, the marks reliability is the content reliability attenuated by examiner unreliability. Thus in the practical situation, it is at the level of marks reliability that traditional and objective test reliabilities should be compared

2 *"Correct mark" and "true mark"*

It would seem wise to restrict the use of the term "true mark" to a meaning which is exactly parallel to the psychometric term 'true score" The "true mark" would then be the mark which the candidate would have received from a perfectly reliable examination, marked by a perfectly reliable examiner When we are talking only about examiner reliability, and not about examination or content reliability, it would

seem wiser to adopt Gulliksen's (1950) term "correct mark" * The "correct mark" is the mark which the answer book should have received if this (unreliable) examination were marked by a perfectly reliable examiner Alternatively, the "correct mark" is the mean of the marks awarded by an infinite number of examiners to the same answer book

It will be noted that these two terms, "correct" and "true' mark, relate to examiner and to marks reliability We have not found any parallel term to relate to content reliability Perhaps one is not needed

3 *Standard error "of marking" "of content" and "of marks"*

The standard error of any examination is (theoretically) the standard deviation of the differences between "true' and "obtained" marks or scores Thorndike (1949) calls this an "absolute" measure of reliability as, unlike the coefficient of reliability (a "relative" measure), it is generally unaffected by differences in group variability For this reason it is "more meaningful if a descriptive statement of the accuracy of measurement is desired" (Thorndike, 1949, p 69)

Standard error of marking This is Taylor's term (1964, p 27) for the standard deviation of the differences between the mark actually awarded by a particular examiner, and the "correct mark" which should have been awarded if his examining were perfectly reliable ** It tells us the extent to which a particular examinee's mark may differ from the mark which he would have received if examiner reliability were 1 00 (Note that for objective examinations the standard error of marking is zero)

Standard error of content For the objective examination, this is the traditional standard error of measurement or SEM For traditional-type examinations, this would be the standard deviation of the differences between the marks which would have been received by the examinees if the examiners were perfectly reliable but the content was not, and those that would have been received if both were perfectly reliable This does not seem to be a very useful concept

Standard error of marks This is the concept which, for traditional type examinations, carries exactly the same meaning as the SEM does for

*Taylor et al (1966 b) have used the term ' true mark ' for this concept Had they differentiated between the three types of reliability, and between ' true ' and correct mark, they might not have been led to be quite so enthusiastic about the 'reliability ' of essay type examinations as their paper suggests

**Taylor et al (1966 b) have used both the terms standard error of marking (page 3) and standard error of estimation (page 6) While the second term is acceptable the first seems preferable as it is unambiguous and carries instant meaning "Standard error of estimation" may be confused by some with standard error of estimate , which is a statistically different concept, with r^2 rather than r in a formula which otherwise is identical

objective examinations Therefore we have often called it by the more familiar term, standard error of measurement This is the standard deviation of the differences between the final mark actually received by the examinee, and the "true mark" which would have been awarded him if both examiner and content reliabilities were 1 00 It tells us the extent to which a particular examinee's mark may differ from the mark which he would have received if examiner and content reliabilities were perfect For traditional-type examinations, this is based on marks reliability For objective examinations, as we have already pointed out, it relates more accurately to the concepts of content reliability , however, since content and marks reliability are identical for objective examinations, there should be no numerical difference in the two SEM's

4 Effect of means and standard deviations

Unfortunately, the above concepts apply strictly only to marks with the same means and standard deviations—a situation which does not exist in India Examiner reliabilities, particularly, can be misleading A coefficient of reliability, of course, is not affected by differences in mean and standard deviation of the two variables correlated Thus the coefficient is quite misleading unless the examinations are scaled to the same standard One examiner could fail all the students and another give them all First Divisions—yet, if they agreed only on the order of merit, examiner reliability would be 1 00

Three simple examples can make this clear

Examiners		Examiners		Examiners	
A	B	C	D	E	F
20	12	20	12	20	20
15	11	19	11	15	15
10	10	18	10	10	10
5	9	17	9	5	5
0	8	16	8	0	0

Note that all three "examiner reliabilities" are 1 00 Yet, without scaling it is obvious that the first and second would not be called "reliable" by most people In the first, the means are the same, but the standard deviations differ widely (ranges 0-20 vs 8-12) In the second, the standard deviations are identical, but the means differ widely (18 vs 10) Only in the third do both means and variances, as well as rank orders, agree Many actual examples will be found in later chapters where apparently high self or inter examiner correlations tend to mask wide difference in M and σ which, in turn, produce large differences in marks for a large number of candidates

The senior author spent half a year searching the literature and

questioning all sorts of people, to try to find an examiner reliability coecffiient which would reflect these differences. Up to the time when the computer work started, none was found. It was only much later that we became acquainted with Buros' unpublished paper (1963), which *does* include differences in means and variances in the formulae defining reliability.

The concept of content reliability does not seem to suffer from this defect, as we tend to think of it as an abstract measure of relative reliability quite unrelated to actual marks. However, when we come to marks reliability, we must again remember that the actual situation, with unscaled marks as they are used in India, is a lower practical level of reliability than the coefficients might seem to indicate.

This leads us to a further differentiation in our terminology. When we are dealing with what Thorndike (1949) calls "relative" measures of reliability—that is, the various coefficients—differences in mean and standard deviation have no effect. However, when we come to the "absolute" measures—the various errors of measurement—these differences make the actual unreliability of marks much larger than it otherwise might be. To avoid misinterpretation, let us add qualifiers to some of the statistical terms that we have already discussed.

5. *Standard errors "(scaled)" and "(unscaled)"*

The standard error of marking and the standard error of measurement that we have so far discussed apply only to the specific scale used by a particular examiner. They will tell us something about inter-examiner reliability only when both examiners are using the same scale. (Or, of course, when the scale used by each examiner is converted to some common standard.) Thus we must call this the "standard error of marking (scaled)". (A more accurate description might have been "standard error of marking (equivalent scales)" but this seemed unnecessarily cumbersome.) It should be obvious that, when examinations are not scaled to a common standard (e.g. those in this study) there is a separate standard error of marking (scaled) for each examiner.

Note that when we talk of the standard error of marking (scaled) we do not necessarily imply that the examiner's marks have been scaled to some common standard. They may or they may not have been. All we are stating is that this is the standard error of marking *for the scale used by this particular examiner*, whatever it may be. Where marks are unscaled, this means that there is a separate standard error of marking (scaled) for each examiner involved.

But when the marks of all examiners are not scaled to the same standard, a list of the standard errors of marking (scaled) of each can

give us a false sense of confidence in the examination It would be quite possible, for example, for the work of a pair of examiners to have an examiner reliability of 90 and standard errors of marking (scaled) of 1 5 and 3 6 respectively—and yet have one examiner pass and the other fail all of the candidates In other words this standard error does not reflect the errors arising from *differences* between the examiners' scales If we consider the "correct mark" for an examinee to be the mean mark of an infinite number of examiners *with the usual range of differences in scales*, then the traditional standard error—which we have called the standard error of marking (scaled)—will not tell us the true extent of variations between obtained and correct marks

Thus we must introduce yet another concept Standard error of marking (unscaled) (Again, standard error of marking (non-equivalent scales) would be more descriptive, but too cumbersome) This is the actual standard error of marking between a pair of examiners using different scales (or even between the two markings of the same examiner, if his scale should change between times)

The difference between these two standard errors of marking can be clearly indicated from the example given in the table in Chapter III, Section 7. For these two examiners (let us call them X and Y here), the statistics are as follows

$$M_X = 8\,1 \qquad M_Y = 22\,2$$
$$\sigma_X = 4\,66 \qquad \sigma_Y = 6\,52$$
$$r_{XY} = 83$$
$$SE_{mkr} = 1\,91 \qquad SE_{mkd} = 2\,67$$

Yet, as can be seen from the table, the mean difference between the two sets of marks is 14 1, and there is *no* difference of less than 8 marks out of 50 Thus, in this particular case, the minimum actual difference is larger than the maximum difference predicted from either of the standard errors While the two standard errors of marking (scaled) are 1 91 and 2 67 the standard error of marking (unscaled), i e the standard error of the difference, is 2 62

6 Mean errors '(scaled) and '(unscaled)

Since the average difference between scores in our example above is 14 1, it can be seen that the standard error of marking (unscaled), which is only 2 62, has not really solved the whole problem This is because it reflects only those errors which are due to differences in the standard deviations of the scales of the two examiners It ignores the difference between the means If the standard deviation differences are large and the means equal then the standard error of marking

(unscaled) will give us useful information. Otherwise, this ignoring of the difference between the means is quite crucial.

The concept of a standard error is that of the distribution of errors around a mean. In the case of the standard error of marking (scaled) *this mean error is zero.* This is why the standard error of marking is an adequate measure, alone, of the error factor in scaled marks—or between examiners using the same scale.

But in case the two examiners differ in mean, as well as in standard deviation, we are faced with yet another problem. The standard error shows the extent to which a student's obtained mark may differ from his "correct mark" or "true mark" (depending on which standard error you are talking about). But when two examiners differ in mean as well as standard deviation, then the mean of each examiner's marks also differs from the mean "correct mark". Thus a constant must be added to (or subtracted from) each examiner's mark to reflect this difference. Where we have a large number of examiners, their average is taken to be the "correct mark". Where we have only a pair of examiners, we would seem to have no choice but to take the average of the two as the best approximation that we have of the "correct mark". The mean "correct mark", therefore, will be half-way between the means of the two examiners—and each of them will differ from it by half of the difference between them. In our example, the difference between the two means is 14.1. Thus the "mean error (unscaled)" is 7.05. (The mean error (scaled), of course, is always zero.)

We should also remember, of course, that this mean has its own standard error, which is related to the formula for the standard error of a difference between means. More about this later.

7. *Range of marking errors (unscaled)*

Can we put these various concepts together? The constant error in each student's mark is plus or minus the mean error (unscaled)—in our example, plus or minus 7.05. But around each end of this range of constant error of 14.1, is a normal distribution of random errors, described by the standard error of marking (unscaled). This leads us to the interesting conclusion that for *some* students, a negative constant error (mean error of marking) may happen to be combined with a large positive random error (SE_{mks}). In such cases (though they would be few) the error component in the unscaled marks may actually be *less* than it would have been if the marks were scaled.

Obviously, the chances of the random component being positive are exactly equal to the chances of its being negative. When we interpret the standard error of marking (scaled) we say that there are two chances

out of three that a particular student s obtained mark is within one SE_{mks} above or below his correct mark A reasonable interpretation of the range of marking errors (unscaled) would seem to be this There are approximately two chances out of three* that a particular student s obtained mark is within a range indicated by the correct mark plus the mean error (unscaled) plus one standard error of marking (unscaled) and the correct mark minus the mean error (unscaled) minus one standard error of marking (unscaled)

When (as is usually the case) there are more than two examiners involved then the constant itself becomes variable It may be reasonable in such cases to base the range of marking errors (unscaled) on the *maximum* difference found between any pair of examiners This is the only realistic description of the hazards of the exam nation

Finally the above statements have revolved around the standard error of marking Somewhat similar statements can be made substituting the standard error of marks or measurement for the SE_{mks} The situation is further complicated by the fact that the content reliabilit es of the two examiners may also differ However these differences are reflected in the SEM (scaled) for each examiner

8 *Applications*

The applicabil ty of these various concepts may be summarized as follows

Means	Std Devs	Concept
Same	Same	Standard Error of Mark ng (scaled)
Same	D fferent	Standard Error of Mark ng (scaled)
D fferent	D fferent	±Mean error (unscaled) ±SE_{mks} (unscaled)

The second is mentioned only because in some universities in India scaling means only adding or subtracting a constant to make all the means equal Since differences in mean are generally (though not always) larger than differences in standard deviat on this pract ce is at least a step in the right d rection

9 *Determination a id formulas*

Table 7 1 summarizes the various measurement models and their related coeff cients and statistics Once the various reliab lit es (examiner content and marks) are obtained the defin tions of correct and true score and of the various standard errors follows

*At the pos tive end of the range of constant error 1/6 of the random error s above + 1 SE_{mks} and 5/6 below it At the negat ve end 1/6 is below − 1 SE_{mks} and 5/6 above it Multiplying the probab l t es $5/6 \times 5/6 = 25/36$ which s very close to 2/3

Examiner reliability: This, as we have seen, is simply the product-moment correlation between the pairs of marks awarded the same answer books, either by the same examiner marking twice, or by two different examiners. If it is a self-consistency coefficient, the two markings must be far enough apart so that memory does not play a contaminating role.

Unfortunately, there seems to be no way to estimate examiner reliability in the most common situation, i.e. when a single examiner marks a single examination (i.e. a single set of answer books).

Where there are parallel examinations, with one examiner per form, examiner reliability, though not determinable, is not really needed. Marks reliability can be calculated without it.

Content reliability: The best way to determine content reliability is to use two examinations that are supposed to cover the same course (say, the Board examinations of two different years). Both of these examinations must then be administered to the same group of students, and the marks correlated. This is very seldom done. So far as we know, the study of this type now being completed by the junior author (Misra, 1970) is the only one of its type for Indian examinations.

Gulliksen (1950) provides the formula for content reliability, under the above experimental conditions. Simplifying his notation somewhat, we may write:

$$r_{cc} = \frac{r_{AB}}{\sqrt{r_{12}\ r_{21}}}. \qquad (7.2)$$

where: r_{cc} is the content reliability of the examination
r_{AB} is the marks reliability of the examination
r_{12} and r_{21} are the two coefficients of examiner reliability.

There are, of course, four estimates of r_{AB}, depending on which pair of different-form examiners one selects. In practice, therefore r_{AB} is taken as the mean of r_{13}, r_{14}, r_{23}, and r_{24}.

In the present study, however, we did not have parallel examinations. For objective-type examinations, when parallel forms are not available (or cannot be administered), we have convenient substitutes: either the split-half or Kuder-Richardson formulas. These, in effect, treat each individual question as a separate examination, and then find out if they are homogeneous enough to consider the total score reliable

This method has been adapated to essay-type examinations through the analysis of variance (Hoyt & Stunkard 1952, Ebel 1965. Ebel's intra-class correlation is also discussed in Guilford, 1954, and Guilford, 1956). Unfortunately, these formulas work only when all candidates answer the same questions. For the situation where a choice of questions is allowed, the senior author has had to develop his own techniques

(Harper, 1966 abstracted in Appendix) Briefly, these are adaptations of Ebel's intra-class correlation The adaptation is based on the assumption that the examiner who allows a choice among a given set of questions is thereby stating that he considers those questions to be of equal difficulty, or at least that equal marks will be received by candidates of equal ability We are, therefore, justified in treating any variance arising from differences in difficulty, as error variance We subtract error variance from the variance of the marks, and divide by the variance of the marks, to obtain our coefficient of reliability The formulas are slightly different for various types of examinations e g "choose five out of nine", "choose three out of six in Part One, two out of four in Part Two ", etc , require slightly different analyses It is well known that allowing a choice of questions does, in itself, reduce reliability (See Harper 1962, DEPSE, 1963, and later chapters of this study)

A reliability formula based on only a single marking of essay type examinations cannot, of course, distinguish between the 'actual values of the questions themselves, and a "halo effect" in marking. By "halo effect" we mean the tendency to allow the mark awarded the first question to influence the marking of the next question However, the *low* content reliabilities reported in this paper suggest that there is little or no "halo" effect in marking This is the more remarkable, since Indian examiners generally read through an entire answer book, marking all questions at once, rather than marking question 1 in all answer books then Question 2 in all answer books, etc The traditional method would seem to discourage completely independent marking of individual questions

In Table 7.1 the content reliability based on a single administration is called r_{kk} to distinguish it from $r_{\alpha\alpha}$ which is based on parallel forms Note, also, that there are as many r_{kk}'s as there are examiners Unless there is a very rigid marking scheme set up, and examiners are thoroughly trained in its use (as is done, for example, in the Advanced Placement essay-type examinations of the College Entrance Examination Board of the U S A), examiners are bound to be looking for different types of "ideal" answers, and to weight different factors differently Thus, the "contents" of a traditional type examination actually varies from examiner to examiner (In one of the subjects in this study, content reliabilities of the same examination ranged from 21 to 76 for the ten examiners.) This, of course is one of the traditional criticisms leveled by opponents of the essay type examination This study provides proof for their contention

In general practice with objective tests and examinations, $r_{\alpha\alpha}$— the parallel forms reliability—is taken as the standard The internal consistency coefficients—r_{kk} in our notation—have been developed

The reader can see from Table 7.1 how formula 7.5 and the traditional formula are applied to the various experimental models we have analyzed.

Standard error of marks (unscaled): This is the most crucial concept, the most needed statistic for application to examinations as they are now administered and marked in India. But we have as yet found no way in which this can be calculated.

Range of total error : For scaled marks, the range of total error is completely defined by the $SE_{m_{kl}}$ (scaled) or the *SEM*. When marks are unscaled, then two components enter in. One is the effect of the combined distributions of errors of the two examiners, which is related to the differences in standard deviations of their individual scales. The other is the constant effect of the differences between the means of their individual scales.

Thus the total range of error in marking for a pair of examiners is found by combining the mean difference between them with the $SE_{m_{kl}}$ (unscaled) for the pair. Actually, the mean of the correct marks is assumed to be half-way between the means of the two examiners. Thus the difference between their means must be divided by 2. This, then, gives the formula which you will find in Table 7.1, in the second box in the row for Range of Total Error.

For the situation where there is a multiplicity of examiners, no exact solution has yet been found. There is little difficulty, of course, if all the examiners have marked the same batch of answer books, or strictly parallel samples (as is done through "randomization" at Gauhati) More often, however, (as in our second study) we have only the results of pairs of examiners on different batches of answer books. The use of the formula for the standard error of a mean would assume that each of these means is a random sample of a set of means. Yet the actual differences found between the means of various examiners are far too wide to support such an assumption.

And for the far more important problem of the range of total error of marks (rather than of marking, as we have been discussing above), we again have found no solution. A lower bound, of course, would be this, based on the widest difference in means found :

$$\pm \frac{M_1 - M_2}{2} \pm SEM_{m_{kl}} \qquad (7.6)$$

But it would be preferable if we could find out the effect on the differences between the means, of the content as well as of the examiner components of reliability.

10. Relationship to the studies in this book

Two studies are reported in this book. One of them, involving 90

for the situation where parallel tests are not available, and derive their value from the assumption that they are rough estimates of r_{cc}. Empirical studies have shown that r_{cc} is generally lower than r_{kk}, probably because of time effects and because of a greater play of sampling error. The junior author has recently completed the first empirical check ever made of the relationship between Harper's r_{kk} and two measures of r_{cc} (Gulliksen's and Lord's), for essay-type examinations (Misra, 1970). He found that the relationship between the two for traditional type examinations is approximately the same as that between Kuder-Richardson reliability and parallel forms reliability, in objective examinations.

Marks reliability. For the parallel examination situation, there are four estimates of marks reliability r_{13}, r_{14} r_{23} r_{24}. For each of these, the marks of two different examiners of two different forms are correlated. For a single representative coefficient, the mean may be taken. In many situations however, it may be more realistic to take the lowest value. It is not the mean but the lowest value which tells what damage can be done. The mean can give us a misleading sense of confidence—like the man who started to climb the Empire State building because he had heard that the average height of buildings in New York was only three hundred feet.

When we have only a single examination marked by two examiners, and an internal consistency coefficient of content reliability, the situation is the reverse of the above. Gulliksen's formula (Formula 7.2) is now rewritten in the following form:

$$r_{AA} = r_{kk} \quad r_{12} \tag{7.3}$$

It will be noted that we have substituted r_{AA} the estimated marks reliability for a single form, for r_{AB} the actual correlation for parallel forms. We have also made the transformation $\sqrt{r_{12}\ r_{12}} = r_{12}$. Either r_{11} or r_{12} can be used, whichever is available, though r_{12} is preferable. There is, of course, a separate coefficient of marks reliability for each examiner, unless the content reliabilities for both of them happen to be identical.

Again, unfortunately, in the *most* common situation—a single examination, marked by a single examiner—we cannot calculate marks reliability. This is because the examiner reliability figure is not available. In a large public examination it may pay to have a representative sample of a hundred or more answer books re marked. They could be sent in batches of, say, 20 each to all of the examiners involved. Re marking just 20 answer books would not take much time. Though individual coefficients would not be reliable (and, in fact, need not be calculated), a coefficient from the pooled data would be a fair estimate of examiner reliability. (However, see discussion of Fig. 10.1.)

If even this cannot be done, it may be better to "guestimate" the coefficient of examiner reliability, rather than to have to do without estimate at all of marks reliability The 'guestimate" can be based on the published literature, including the present study For determining say, the standard error of measurement for use in a selection situation even a "guestimate" would result in a far more accurate interpretation of the examination marks involved than the usual practice of treating them as perfectly reliable

Standard errors of marking and of marks (scaled) As these are all based on the standard SEM formula, no detailed discussion is needed As will be noted from Table 7 1, the standard errors for unscaled marks are different for each examiner involved (Remember that we use the SE_{mks} (scaled), because with unscaled marks, each examiner has his own scale Thus we are describing the standard error in terms of *his* scale) If marks are scaled to a common mean and standard deviation—as they should be—then, as with objective examinations, there is only a single standard error to represent the work of all examiners

Standard error of marking (unscaled) It is when we come to applying the absolute measures to the situation where we have many examiners, each with his own individual scale, and we want to know the error introduced by this fact, that we have to tread on new ground We have assumed that when we have two examiners for the same set of answer books, the best estimate we have of the "correct mark" of each examinee is the mean of these two examiners We may then assume that the difference between a pair of marks is made up entirely of two error components, one contributed by each examiner, plus a constant which is the difference between the means of the two examiners Since the standard deviation is not affected by the mean, we may ignore this constant, and concentrate on the error components In this circumstance, Gulliksen points out that if we can assume that the error variances of the two tests (in our case, we may substitute "examiners" for "tests") is equal, and that the errors correlate zero, then (adapting Gulliksen's formula 4 to our notation)

$$2\,SEM^2 = \sigma_d^2 \qquad (7\,4)$$

where σ_d is the standard deviation of the difference scores i e of the array of marks obtained by substracting one examiner's mark from the other for each answer book By simple transformation we therefore obtain

$$SE_{mks}\ (\text{unscaled}) = \sqrt{\frac{\sigma_d^2}{2}} \qquad (7\,5)$$

when the standard deviations of the scales used by the two examiners are different, formula 7.5 provides a sort of average standard error for the pair

examiners for the same batch of answer books, required some specialized coefficients which will be discussed in the report of that study The second study followed the "single examination, two examiners' model, with both self-consistency and inter examiner coefficients The remaining "cells" in Table 7 1 are not directly relevant to these two studies They are included in the table for completeness, and also to add perspective to the coefficients and measures used

We cannot end this section without acknowledging, again, the existence of a much more complete, and probably much better set of concepts developed by Oscar K Buros (1963) Had these come to our attention at the beginning, instead of after all the computer work was finished, this entire project might have been cast in a different mould Any future researcher in this field should certainly be aware of Buros' conceptualizations before finalizing his research design

11 *A final note*

Several of the concepts in this chapter are still quite tentative, and are based on insights which—though sought after for several years—developed in the mind of the senior author only near the very end of the project They are included here to stimulate thinking and discussion, even though they have neither been tested yet in practice, nor are they reflected fully in the chapters which follow This may be frustrating to some readers The alternative of not including these tentative concepts here, seemed even less desirable

8

*"Ninety Marking Ten"—Detailed Report**

THE BASIC outline of this study has already been given in Chapter III and the experimental design elaborated in Chapter VI(B)

In 1966 Taylor Tluanga and Misra published *A Study in Multiple Marking* Since this is the only other study of its kind done in India, we will refer to it several times in relation to our own findings The general similarity of the findings is the more striking when seen in terms of the differences in experimental design between the two studies These differences may be summarized as follows

	Harper et al	*Taylor et al*
Subject	History	English
Level	Higher Secondary	B A.
Maximum marks	50	100
Number of answer books	10	100
Method of reproduction	facsimile	typing/cyclostyling
Number of examiners	90 from same State and Board	19 from 9 States 18 colleges

However, the parallels are limited by a striking difference which was probably more due to practical limitations in data handling facilities, than to any theoretical considerations Taylor et al studied 19 examiners and 100 answer books, so their analysis deals primarily with examiners,

*Parts of this chapter have already been published (Harper 1967) but this chapter contains several important analyses that did not find a place in the summary article

we studied 90 examiners and 10 answer books, so our analysis deals primarily with answer books Thus the two studies complement each other, and only occasionally can they be compared directly

It is interesting to note that both the present study and Taylor et al s were carried out within the same two year span, and that the same junior author was involved in both of them

1 Frequency distributions of marks

Table 8 1 gives the frequency distributions of the marks awarded each of the ten candidates (A to J) These are marks out of 50 for a *History examination* In the first 10 columns the results indicate the frequency with which each mark appeared when answer book A was marked first, B marked second, etc The second set of 10 columns indicates the frequencies when the marking was in reverse order, i e answer book J, then I, then H, etc Table 3 1 (Chapter III) combined the distributions for the two different orders into a single table

By inspection of the two sets of columns it can be noted that the distributions for each answer book are roughly similar, for the two orders of marking Neither the means nor the medians differed by *more than two or three points* These differences are treated in more detail in section 12 on 'Position Effects'

We have already commented on the fact that the sums of the two orders of presentation (columns A J and J A) show nearly identical distributions, thus establishing our confidence in the total sample

We have already noted also that the highest mark awarded any candidate by any examiner is 38 This is 76% marks which, in this state, means the award of a "Distinction" to Candidate E by one examiner Note that this same candidate was also failed (less than 33% marks, i e 16 or lower) by seven examiners

The lowest mark awarded was 2, or 4% One of the candidates receiving this extremely low failing mark was also passed by two of the 90 examiners

The widest distribution of total marks is for answer book E (see *Table 3 1), a range* of 28 marks* The narrowest distribution (Candidate B) is one half of this, i e 13 This still represents a *minimum* difference, for this study, of 26% marks between the highest and lowest awarded a single candidate Perhaps teachers and students should be most concerned with the maximum range (56%) as this defines the risk they

*In this paper the convention of defining the range as including the highest and lowest mark has been followed Thus if the highest mark is 3 and the lowest is 1, this is a range of 3 Some writers take only the difference as the range, and would call this a range of 2

take when students sit for an examination Thus a candidate in a Board Examination in class X History may get either a Distinction or a bad Fail, depending purely on "luck" (*i e. who his examiner happens to be*) —and regardless of his personal "merit".

How often might this happen ? The correct mark (see Chapter VII) or average given by these 90 experienced examiners to this particular answer book was 23.13 (cf. Tables 3.1 and 8 6). It is not unreasonable to assume that, in a large examination, say 1500 "should" receive a mark of 23. We can also assume that these candidates are roughly similar to Candidate E, and the examiners are roughly similar to our ninety. As many as 17 of these 1500 candidates may receive a Distinction, while about 117 of them will fail—even though they are actually of *equal* merit.

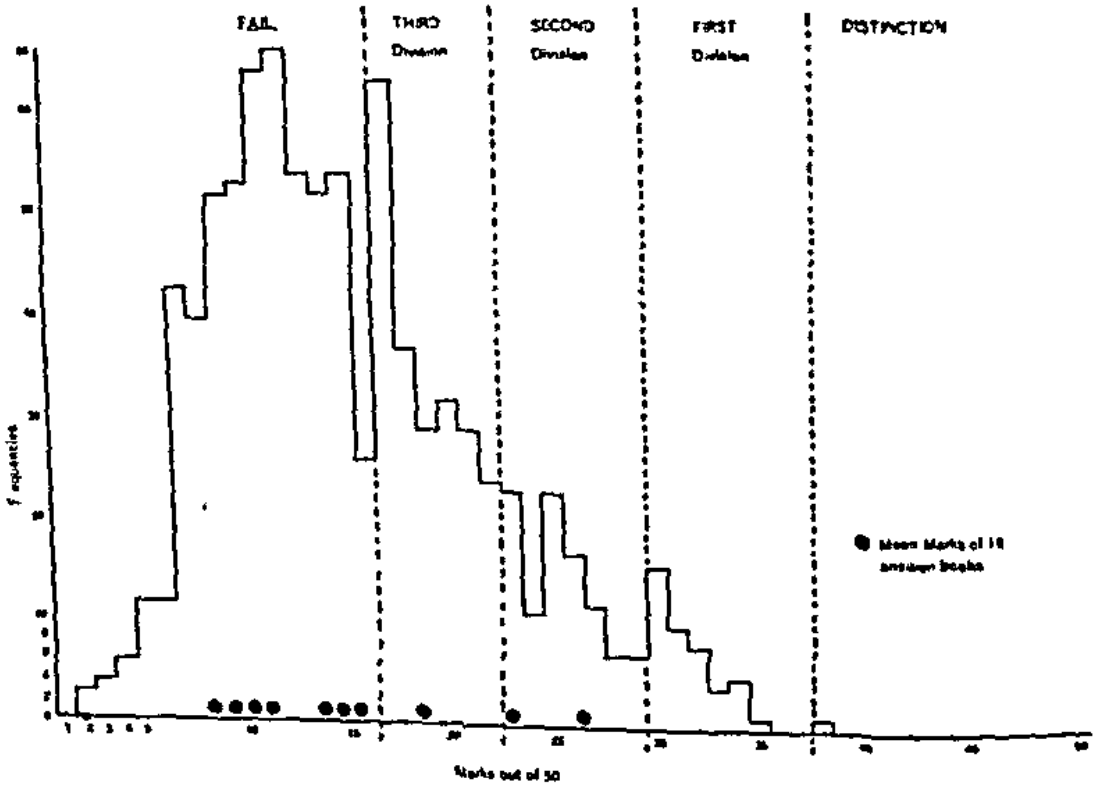


FIG. 8 1

Figure 8.1 presents the total frequency distribution graphically. Note that this is not the distribution of marks as desired by any one examiner. It is the combined distribution of ninety examiners, with different means and standard deviations. Thus, it is a much wider distribution than that of any single examiner would be. Remember, however, that distributions of public examination marks are also widened by the same effect—the pooling together of the marks of many examiners, each with

FREQUENCY OF [illegible] EXAMINERS) AND TOTAL (90 EXAMINERS)

Marks	and (N=90)						TOTALS		
	A	B	G	H	I	J	A J	J A	G T
38								1	1
37								—	—
36								—	—
35								1	1
34							2	3	5
33							3	1	4
32							5	3	8
31							4	6	10
30							7	9	16
29							3	4	7
28							1	6	7
27							5	7	12
26					1		6	11	17
25					1		13	10	23
24					—		6	5	11
23					—		10	13	23
22					3		10	14	24
21	2				1		16	13	29
20	1				6		14	18	32
19	—			1	2		13	16	29
18	6			2	7		18	19	37
17	6	1	2	4	18		30	33	63
16	3	1	1	3	3		12	14	26
15	4	3	2	7	14	3	26	28	54
14	4	6	5	3	11	4	23	29	52
13	5	6	13	7	2	5	26	28	54
12	2	5	13	9	11	6	29	37	66
11	3	6	8	12	4	8	31	33	64
10	3	7	14	9	2	10	24	29	53
9	2	2	9	14	1	8	28	24	52
8	—	3	13	7	1	14	12	28	40
7	1		3	4	2	13	20	23	43
6			2	1		7	6	6	12
5			2	3		6	7	5	12
4			—	3		1	5	1	6
3			1	1		3	4	—	4
2						2	1	2	3
			90	90	90	90	420	480	900

NOTE In or range of a d stribution

his own individual "standard" of marking

It is particularly interesting in Figure 8 1 to note H J Taylor s "J-effect" operating Examiners tend to avoid the mark immediately below a particular Division, particularly Third Thus there are 63 at 34%, only 26 at 32%, but again 54 at 30% An experienced examiner, asked to comment on this writes 'The examiner does not, usually, want to fail a student by one mark and so if he finds that the total marks come to 16 he usually adds one half mark or two half marks to questions to make the total 17 However, when a student gets a total of 15 the examiner usually is not inclined to give two extra marks to pass him That is why you have got a pile up of marks at 15 and 17 and not at 16'

The same is true for First Division It is interesting, though that that there is no J-effect at the division between Second and Third Perhaps this is because the differentiation between Second and Third does not carry nearly the consequences that First' and Fail do

There is, however, apparently a J-effect at 25 marks right in the middle of Class Two—and a similar, though smaller effect in the middle of Class Three How to explain these? The same experienced examiner quoted above offers this " the only reason I can think of is that when an examiner marks an answer book he probably decides on the basis of the first one or two questions which division the student deserves If he decides that the student deserves Second Division then on an average he gives 5 marks out of 10 for each question because if he gives an average of 6 marks the student gets a First Class (60%), if he gives 4 or 4½ he gets a Third Class " This seems as reasonable an explanation as any Note, also, that 7 seems to be the lowest mark for most examiners

Another interesting fact about the J-effect may be noted from Tables 8 1 and 3 1 The strongest J-effect in each distribution occurs near the average mark for that candidate For example, candidate D s average is 27 marks, his J occurs at First Division (30 marks) The J at Third Division is for answer books I, F, and A, whose averages are 15 and 14 marks The only J s at 23 marks are for D and E, with means of 27 and 23 For answer book C, with a mean mark of 19, there is a sort of reverse J at 22 marks—as though many examiners wanted to give him the top of Division Two, but no higher These J effects are, of course, for pooled group results, but as we shall see later, individual examiners also exhibit a J-effect

All this unevenness makes the scaling of marks *by standard deviation methods* a questionable procedure Such scaled marks would have the same distribution as the raw marks, but with the J s coming at other than the dividing points between Divisions Scaling should be by a normalizing procedure, as we shall see later

2 Classes or Divisions

The lines across Table 8 1 represent Divisions—the lowest being the distinction between passing and failing In many cases it is not the marks themselves, but rather the Divisions (or Classes) that count If we can treat Divisions as roughly accurate, it would be very helpful Can we ? See Tables 3 2 and 8 2 Only one answer book falls definitely in only one Division (Candidate J was failed by all examiners) Another three were in only two Divisions (Even the most accurate existing test will, of course, put some borderline candidates in two Divisions But these three were not borderline) Four candidates were placed in three different Divisions by various examiners and two were placed in all four Divisions

Table 8 2 also makes it evident that the two batches of examiners agreed quite closely in distribution of marks, and that the order in which the answer books were marked made no striking difference in these distributions The large difference for I is due to the fact that many more examiners marked it just at the passing point (17) when it was second than when it was nineth in order of marking Again however, there is no consistent tendency to do this for all answer books

It is obvious from all these tables that, for a particular answer book, even ' Division" has very little meaning in evaluating candidates It should be remembered, however, that when several unreliable but inter-correlated scores or marks are added together (and several Indian studies have shown marks in different subjects to be correlated), the reliability of the sum is raised This follows from the formula for "battery reliability ', which can be found in standard texts Division or Class on the aggregate may, therefore, have considerably more meaning than Division or Class on any single paper *

It is also interesting to note that the range of marks is wider for the better candidates, and lowest for those near the bottom (This is further confirmed by the *s*'s in Table 8 6, and is the opposite of the phenomenon found for rank orders) Thus we may conclude that we can place more reliance on the statement "He is a failure" than on the statement "He is a First Class student," when marks are not scaled

*There seems to have been no study of aggregate marks reliability, as against reliability of separate tests or papers However, one line of evidence comes from the fact that studies done in India have generally shown a much higher validity for predicting aggregate marks than for predicting marks in any single subject There are also some useful indications given in the study of Datta (1967) (This able young scholar died tragically in an air crash It is hoped that the research which he completed just before his death will keep his memory alive) Datta correlated aggregate marks on half yearly and annual examinations for eighth grade in 15 schools in Tripura The coefficients for individual schools ranged from 80 to 93, with a median value of 88

TABLE 82

FREQUENCIES WITH WHICH EACH CLASS OR DIVISION WAS AWARDED BY 90 EXPERIENCED EXAMINERS TO EACH OF 10 HISTORY ANSWER BOOKS, IN THE FIRST AND SECOND ORDER OF MARKING

Answer book	A		B		C		D		E		F		G		H		I		J	
Order	1	10	2	9	3	8	4	7	5	6	6	5	7	4	8	3	9	2	10	1
Division										1										
I					1		16	19	4	4										
II		1			7	8	15	21	18	23	1	1						2		
III	15	15	1	1	26	24	8	8	18	15	17	20	2		3	4	11	26		
I & I																				
20%–33%	24	30	34	30	8	15			2	5	20	24	19	25	23	27	25	13	13	23
0%–19%	3	2	7	17		1					4	1	21	23	16	17	3	1	29	25
Total	42	48	42	48	42	48	42	48	42	48	42	48	42	48	42	48	42	48	42	48

3. Reliability of pass/fail

Even on the crucial distinction between "passing" and "failing", there is very little agreement among these experienced examiners. Only one candidate was passed by all 90 examiners, and only one was failed by all. Eight of the ten candidates were passed by some examiners and failed by others.* These facts suggest that for over three-quarters of the papers in any public examination, we cannot even be sure whether they are really of "passing" or "failing" quality. Thus, whether a particular candidate passes or fails is as much a matter of chance (the chance of who his examiner is) as of anything else.

Though uncertainty may be inevitable at the border, where a change of one mark (16 to 17, or 17 to 16) makes the difference, most of the affected candidates were nowhere near the border. Since failure in a single paper can mean failure in the examination, no matter how high the remaining marks, this is a serious matter indeed. The only solution—other than better and more reliable examination methods—is to apply both randomization and scaling of answer books (Taylor 1963c), and Taylor's (1964) grace marks techniques (see Appendix B of this volume), which take realistic account of examiner error. Dr. Taylor's experience (mentioned in a personal communication) was that when these methods were applied at Gauhati, the pass percentage went up. It is thus obvious that a substantial proportion of the "failures" in our public examinations are failures of the system, not of the candidates.

The system fails, and we blame the students! Why should any student be labelled a "failure", and his life and finances ruined, when his "failure" is not due to lack of ability on his part but to lack of reliability in the examination—especially when the simple techniques of randomizations, scaling, and a scientific (rather than arbitrary) grace marks system can eliminate most of these false "failures"? These minimum changes require no change in the traditional type of questions, no re-education of examiners, no re-orientation of teaching methods, no re-education of students. They require only a simple administrative decision about the way in which examination answer books, and examiners marks, are to be treated in the Board's office. Is there any reason why this decision cannot be made tomorrow?

4. Means, standard deviations, and pass percentages of examiners

Thus far we have been dicussing variability of marking in terms of the marks awarded each of the ten answer books. The source of this variability, of course, is not the answer books themselves, but the

*Though B and G were failed by two examiners each, only one examiner failed both of them—a typical lack of agreement.

examiners Taken as 90 individuals how do the examiners vary ?

We have already seen in the Highlights chapter (III Section 4 Table 3 3) that the pass percentages varied from 10% to 80% The comparable figures from Taylor et al (1966) are 7% to 62%

Perhaps more precise measures of how examiners differ are the means and standard deviations of the marks they award Table 8 3 gives the frequency distribution of the means of the 90 experienced examiners This shows how widely these examiners all presumably trained to examine at the same standard differ on the average merit of this batch of ten answer books The range is from 8 1 to 22 2 marks out of 50 In other words one experienced examiner thought the average merit of this group of candidates was only 16% of marks—while another thought their average merit was 44% of marks The comparable figures in Taylor et al s study are 19 2% to 35 0% The larger number of answer books (100 instead of 10) tends to make the extremes regress toward the mean , However, even in Taylor et al s study the highest mean is nearly double the lowest

TABLE 8 3

FREQUENCY DISTRIBUTION OF THE MEANS OF THE MARKS AWARDED BY 90 EXPERIENCED EXAMINERS TO THE SAME 10 HISTORY ANSWER BOOKS

Means	Frequencies
22.0—22.9	1
21 0—21.9	—
20 0—20.9	1
19 0—19 9	7
18 0—18 9	9
17 0—17 9	10
16 0—16 9	15
15 0—15 9	10
14 0—14 9	12
13 0—13 9	12
12.0—12.9	4
11 0—11 9	2
10 0—10 9	4
9 0— 9 9	2
8 0— 8 9	1
Total	90

Range of Means = 8 1—22.2

One is forced to speculate that the only reason the state-wide pass percentage remains fairly steady is that the results of a very large number of examiners are thrown together and that it is only this average

of averages and not the standards of the individual examiners) that remain relatively steady from year to year. Sampling theory would support this speculation.

Examiners differ not only in mean marks, but also in the standard deviations of their individual marking systems (It is the complex interrelationship between these two which produces the differences in pass percentage.) Since a standard deviation based on only 10 marks is of limited reliability, the labour required to calculate these did not seem justified by the potential results (since this study was done by hand, and not by computer) Our main consideration (as will be seen later) was the average standard deviation of the marking system for the batch of examiners as a whole. This σ must be defined as the *average* σ of the 90 examiners, rather than the σ of the average marks. For this purpose, a rough estimate seemed to be an adequate substitute for the detailed calculations. A cue was taken from Jenkins' method for estimating σ. Snedecor (1940, p 85) reports (and Guilford (1956) quotes him) that the average range for 10 *measures* is 3.08 times their standard deviation. This, then, gives us a method for estimating σ from the range. The estimation process is given in Table 8.4 which, incidentally, also gives a frequency distribution of ranges and estimated standard deviations for the 90 examiners. While this method may not give a very accurate result for any single examiner, it is reasonable to assume that the average over 90 examiners is a fairly accurate figure. Thus we arrive at a standard deviation of marks of 6.367.* Since these are marks out of 50, the figure yields 12.73%. This agrees closely with other studies done in India, and suggests that this sample of 10 examination books is a representative sample of High School History examination books of this type.

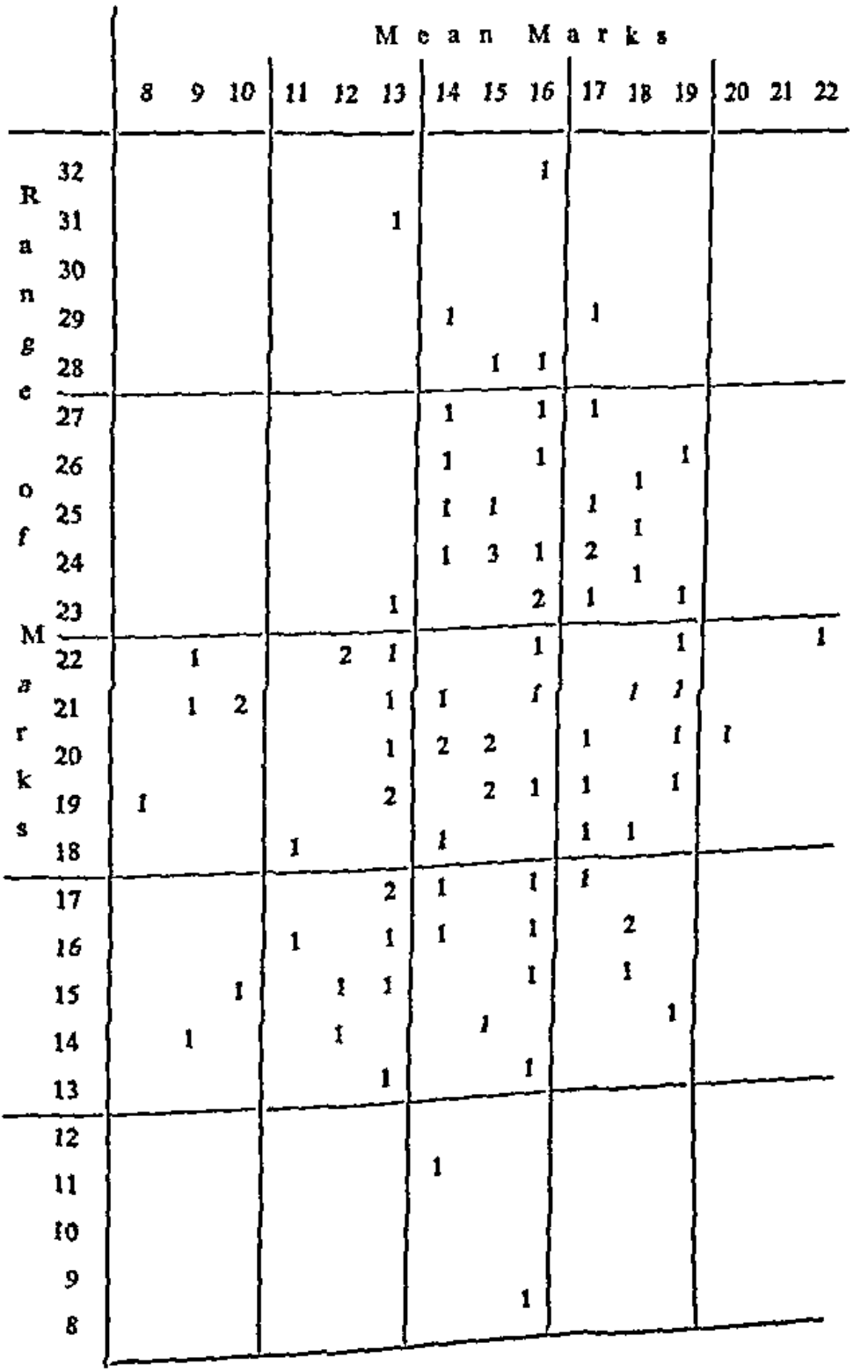
Recognizing the limitations of each individual σ_{m}, we still may compare these results with those of Taylor et al. Translating their "marks out of 100" to "marks out of 50", the comparable range is 3.48 to 5.58 for their nineteen examiners. Their mean σ was equal to 4.54 out of 50, as against ours of 6.67. Whether these differences represent differences between the marking of History and of English, or whether they have some other source, we cannot be certain.

It is clear, however, that both in terms of mean and of standard deviation, each individual examiner seems to have his own particular marking system or scale. Figure 8.2 examines the relationship between range (as an estimate of standard deviation) and mean, for the "scales"

*After these σ's had been estimated, a friend offered to run some of the data through a computer. Because of computer limitations, we had to select a stratified random sample (described later) of 49 examiners' results. For this sample, the mean σ was 6.165, and the range [illegible] to 10 1. These exact calculations from a sample of 49 agree very closely [illegible] exact" estimates from the entire 90.

FIG. 82

Relationship between mean mark and range of marks in the scales used by 90 experienced examiners when marking the same 10 History answer books



of averages (and not the standards of the individual examiners) that remains relatively steady from year to year. Sampling theory would support this speculation.

Examiners differ not only in mean marks, but also in the standard deviations of their individual marking systems. (It is the complex interrelationship between these two which produces the differences in pass percentage.) Since a standard deviation based on only 10 marks is of doubtful reliability, the labour required to calculate these did not seem justified by the potential results (since this study was done by hand, and not by computer). Our main consideration (as will be seen later) was the average standard deviation of the marking system for the batch of examiners as a whole. This σ must be defined as the *average* σ of the 90 examiners, rather than the σ of the average marks. For this purpose, a rough estimate seemed to be an adequate substitute for the detailed calculations. A cue was taken from Jenkins' method for estimating σ. Snedecor (1940, p 85) reports (and Guilford (1956) quotes him) that the average range for 10 measures is 3.08 times their standard deviation. This, then, gives us a method for estimating σ from the range. The estimation process is given in Table 8 4 which, incidentally, also gives a frequency distribution of ranges and estimated standard deviations for the 90 examiners. While this method may not give a very accurate result for any single examiner, it is reasonable to assume that the average over 90 examiners is a fairly accurate figure. Thus we arrive at a standard deviation of marks of 6.367.* Since these are marks out of 50, the figure yields 12.73%. This agrees closely with other studies done in India, and suggests that this sample of 10 examination books is a representative sample of High School History examination books of this type.

Recognizing the limitations of each individual $\sigma_{est.}$ we still may compare these results with those of Taylor et al. Translating their "marks out of 100" to "marks out of 50", the comparable range is 3.48 to 5.58 for their nineteen examiners. Their mean σ was equal to 4.54 out of 50, as against ours of 6.67. Whether these differences represent differences between the marking of History and of English, or whether they have some other source, we cannot be certain.

It is clear, however, that both in terms of mean and of standard deviation, each individual examiner seems to have his own particular marking system or scale. Figure 8.2 examines the relationship between range (as an estimate of standard deviation) and mean, for the "scales"

*After these σ's had been estimated, a friend offered to run some of the data through a computer. Because of computer limitations, we had to select a stratified random sample (described later) of 49 examiners' results. For this sample, the mean σ was 6.165, and the range 2.7 to 10.1. These exact calculations from a sample of 49 agree very closely with our "inexact" estimates from the entire 90.

FIG 8 2

Relationship between mean mark and range of marks in the scales used by 90 experienced examiners when marking the same 10 History answer books

Range of Marks	Mean Marks 8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
32									1						
31						1									
30															
29							1			1					
28								1	1						
27							1		1	1					
26							1		1			1			
25							1	1		1	1				
24							1	3	1	2	1				
23						1			2	1		1			
22		1			2	1			1			1			1
21		1	2			1	1		1		1	1			
20						1	2	2		1		1	1		
19	1					2		2	1	1		1			
18				1			1			1	1				
17						2	1		1	1					
16				1		1	1		1		2				
15			1		1	1			1		1				
14		1			1			1				1			
13						1			1						
12							1								
11															
10															
9															
8									1						

of the 90 experienced examiners There seems to be virtually no relationship

Thus the candidate in a large public examination with, say, one hundred examiners, is faced with one hundred different marking scales on the basis of which his "merit' may be reported Not only does this

TABLE 8 4

ESTIMATION OF σ FROM RANGE FOR 90 EXAMINERS, TO FIND MEAN σ OF MARKING SCALE USED

Range*	R 1*	f	σ_{est}
8	7	1	2.3
—			
11	10	1	3.2
—			
13	12	2	3.9
14	13	4	4 2
15	14	5	4 6
16	15	6	4 9
17	16	5	5.2
18	17	4	5 5
19	18	8	5 8
20	19	8	6.2
21	20	8	6 5
22	21	7	6 8
23	22	6	7 1
24	23	8	7 5
25	24	4	7 8
26	25	4	8 1
27	26	3	8 4
28	27	2	8 8
29	28	2	9 1
30	29	—	—
31	30	1	9 7
32	31	1	10 1
		90	

σ_{est} = estimated standard deviation

$$= \frac{1}{3\,08} \times \frac{\Sigma f(R-1)}{N}$$

$$= \frac{1765}{(3\,08)\,(90)}$$

$$= 6.367$$

*We have followed the convention of defining the range as *including* the highest and lowest mark Snedecor has defined the range as the *difference* between the highest and lowest Therefore one has had to be subtracted from these ranges in order to make them equivalent to the ones on which Snedecor based his table

destroy an assumed comparison between different candidates (unless they happen to have the same examiner)—it is obvious that sheer chance (i e the chance factor of who his examiner may be) is likely to be far more important than the individual merit of the particular student, in determining what Division he will be placed in, and even whether he will pass or fail

5 Frequency distribution of rank orders

Differences in mean and standard deviation among examiners can be handled with simple statistical procedures Just as we can convert degrees Fahrenheit to degrees Centigrade, so also can we convert the individual scales of all possible examiners to a single agreed upon common scale The formula for one method for doing this has already been given in Chapter VII (formula 7 1)

The use of formula 7 1 (or any other method) will completely wipe out the inter examiner differences shown in Tables 8 3 and 8 4, and Figure 8 2 But will it also eliminate the wide distribution of marks shown in Tables 3 1, 8 1 and 8 2 ? The answer is yes *if* examiners can at least agree on the rank order of each candidate We must now examine this "if"

The ten marks awarded by each examiner were converted into rank orders The answer book with the highest mark was given a rank of 1 , that with the lowest mark a rank of ten Tied marks were given the average rank of the ties, as per accepted procedure Tables 3 4 and 8 5 summarize the distributions of rank orders in the same manner as Tables 3 1 and 8 1 summarize the distributions of marks

Table 8 5 gives the distribution of rank orders for each candidate, in both orders of marking The half ranks (1 5, 2 5, etc) are, of course, the results of ties Notice that the distributions for each answer book are approximately the same for first and second order of marking There were small differences in means and medians, almost none of them more than half a rank

Table 3 4 summarized the two parts of Table 8 5, giving the total frequency of each rank for each candidate The ranks were rounded, to make the table easier to read (E g a rank of 2.5 was counted as a 2, since it was the intention of the marker to give these two tied papers the second highest score)

As can easily be seen, there is considerable variation in range of ranks—from 3 to 10 There is a high degree of agreement that candidate D is the best (86% of the 90 examiners put him first, i e rank 1 or 1 5)

It is interesting to note, however, that out of the 77 who considered D the best (rank 1), 32 awarded him First Class marks, 34 Second Class,

and 11 Third Class. The examiners may think him best, but they certainly do not agree on how good he is.

There is somewhat less agreement as to which candidates are worst. But the most disagreement is about the average candidates. Candidate F, with an average rank order of 4.88, is given, by one examiner or another, almost every rank available from 1 to 10. That this is not an unusual result is indicated by the fact that Taylor et al (1966 b) report that one of their answer books was ranked 3rd by one examiner and 98th by another.

Unfortunately, under the Indian system the decisions as to whether a candidate should Pass or Fail are made in the middle of the range. If either a much larger, or a much smaller percentage were allowed to pass, we could be much more certain that "Pass" and "Fail" represent descriptions of actual ability or achievement, rather than just good or bad luck for an average candidate.

We started with the statement that scaling will eliminate differences in marks awarded by different examiners, if those examiners at least agree on "order of merit". We see that they are by no means unanimous. However, as already hinted at in Chapter III, the agreement on ranks is somewhat better than the agreement on marks. Therefore the conversion of each examiner's marks to a common scale will at least reduce the variability in marks We will postpone further discussion of this to the chapter on "Scaling".

6. *Means, medians and standard deviations**

Table 8.6 reports the mean mark, mean rank order, median mark and standard deviation of marks and rank orders, for each of the ten answer books. Table 8 6(a) gives these facts for the two orders of marking separately. Table 8 6(b) gives them for the pooled data. In Table 8.6(b) they are arranged in order of the size of the mean, so as to bring out the relationships among these various statistics. We will concentrate our discussion on these pooled data. (The reader is referred to Chapter VII for a detailed discussion of the terms used.)

Correct mark : The mean of the 90 marks awarded a particular answer book can be considered to be, practically by definition, the "correct mark" for that answer book. The definition, of course, requires an infinite number of examiners, but 90 is a reasonable sample. This makes the standard error 9½ (i.e. the square root of 90) times smaller than the standard error of marking for any individual examiner. The standard errors of the correct marks (not reported in the table) range

*The organization of this particular section owes a deep debt to Taylor et al. (1966 b).

from 0 274 to 0 500 so we may regard these correct marks as quite stable values

The reader is reminded again that the correct mark is not the 'true' mark of the examination The correct mark for each answer book is the mark which would have been awarded if each examiner were perfectly reliable It still contains however the error which is due to the unreliability of the examination itself regardless of the reliability of marking

The mean of the correct marks is 15 52 their range is 8—27 and their standard deviation is 5 965 The comparable figures for Taylor et al s (1966 b) study (if we convert them from marks out of 100 to marks out of 50) are 13 7, 3—24 and 3 96 respectively Whether these d fferences reflect differences in the scales commonly used for history and for Engl sh or whether they are due to sampling differences we do not know

It is instructive to compare the means of the individual examiners (Table 8 3) with the mean of the correct marks Only ten of these experienced examiners had approximately the same mean Thirty seven of them thought the average merit of this batch was lower than 15 and 43 of them thought it was higher Similarly the pass percentage for the correct marks is 30% (Taylor et al found the same pass percentage) But as can be seen from Table 3 3 only 23 of our 90 experienced examiners had the same pass percentage 17 passed fewer and 50 passed a larger percentage of the candidates

Standard deviations in Table 8 6 The columns headed s give the estimated standard deviations in the populations from which these are samples In other words they were derived using N—1 in the denominator of the σ formula This makes very little difference when estimating say, the s for the marks for Candidate D as we are est mat ng on the basis of a sample N=90 However σ for the 10 Mean Marks was 5 659 while s is 5 965 a difference of 5% This is justified as we are estimating the standard deviation of the population of marks of which this is a sample

Standard error of marking The standard deviations of the marks awarded each answer book are of course the standard errors of marking for those answer books These are the standard deviations of the arrays in Table 3 1 and represent the distribution of awarded marks around the mean or correct mark

The range of standard errors of marking is from 2 6 to 4 75 with a mean of 3 66 This mean standard error of marking is the best absolute measure that we have of the reliability of this examination Since Taylor et al (1966 b) calculated their standard errors for examiners, wh le we have calculated them for answer books the two sets of statistics are not directly comparable However the means for the total examiner answer

book experiment are comparable. As against our standard error of marking of 3.66, the comparable figure for Taylor et al. (1966 b) is 2.21. Whether these differences are due to subject matter (history vs. English) or to the nature of the sampling (90:10 vs. 19:100) is not known. However,

TABLE 8.6 (a)

TEN ANSWER BOOKS MARKED BY 90 EXAMINERS. MEANS, MEDIANS AND STANDARD DEVIATIONS OF THE MARKS OF 90 EXPERIENCED EXAMINERS FOR EACH ANSWER BOOK, AND THE MEANS AND STANDARD DEVIATIONS OF THE RANK ORDERS ASSIGNED, FOR EACH ORDER OF MARKING.

	Marks		Ranks
	Mean	σ	Mean
A	14 60	3.393	5 23
B	11.69	2 484	7.33
C	19.64	4.178	3 15
D	26.55	4 910	1.23
E	22.93	4 256	2 12
F	15 52	3 884	4.65
G	9 81	3.094	8.44
H	10 42	3.343	8 01
I	14.24	3.289	5.48
J	8.10	3 260	9 36

	Marks		Ranks
	Mean	σ	Mean
J	9.38	2.780	9.10
I	16 56	3.524	4 36
H	11 40	3 613	7.52
G	9.73	2 648	8.76
F	15 21	4 068	5 04
E	23.32	5.182	2 08
D	27.35	4.250	1.18
C	18.42	4.272	3.44
B	10 83	2 652	7.99
A	14 42	3.512	5.52

TABLE 8 6 (b)

TEN ANSWER BOOKS MARKED BY 90 EXAMINERS MEANS, MEDIANS AND STANDARD DEVIATIONS OF THE MARKS OF 90 EXPERIENCED EXAMINERS FOR EACH ANSWER BOOK, AND THE MEANS AND STANDARD DEVIATIONS OF THE RANK ORDERS ASSIGNED, ARRANGED IN ORDER OF AVERAGE MARKS.

Answer Book	Marks			Rank Order of Mean Marks	Rank Orders	
	Mean	s	Median		Mean	s
D	26 98	4.562	27.1	1	1.2	.43
E	23.13	4 750	23.1	2	2.1	.71
C	18 99	4 249	19 4	3	3 3	1 03
I	15.48	3 592	15.3	4	4 88	1.44
F	15.36	3 964	15 3	5	4.86	1.39
A	14 50	3.439	14 3	6	5.4	1.16
B	11.23	2.597	11.3	7	7 68	1.22
H	10 94	3.504	10.7	8	7.75	1 33
G	9.77	2.848	9.6	9	8 6	1.17
J	8.78	3 064	8.4	10	9.2	.98

For the 10 answer books : M=15 52; Mean s=3 66=standard error of marking

s of Mean Marks=5 965=s of correct mark

For total distribution of Marks (from tables not shown here)

M=15.52; σ=7.187

σ of Examination (see Table 8.4 for method of estimating)=6.367

Note : s is estimate of population σ from sample

several studies have shown the content and/or marks reliability of English examinations to be *higher* than, for example, traditional mathematics examinations (see Chapter V), and just possibly the same factors may be operating to make the marking of English examinations more reliable than those of some other subjects

Average scale used by examiners There is as we have seen (Figure 8 2) a wide variety of scales in use by our 90 experienced examiners all, supposedly, working under the same assumed standard There is, perhaps some advantage in knowing the characteristics of the average scale used

The calculation of the mean of this average scale was quite simple. But to calculate and average 90 standard deviations, in the absence of a computer or electric calculating machine, was a rather large job The method used for estimating the mean standard deviation has already been discussed in relation to Table 8 4

The characteristics of the average marking scale used by these 90 examiners are Mean=15 52, standard deviation=6 367 (The comparable figures for Taylor et al (1966 b) are 13 7 and 4 54) But when individual standard deviations for the same batch of ten answer books range from 3 to 10 (Table 8 4) the significance of the average scale is perhaps minimal

The standard deviation of the correct marks (5 965) is, of course less than the average standard deviation of the scales used This is because the standard deviation of any one examiner s marks includes a random error factor which is absent (or, at least, reduced 9½ times) in the standard deviation of the correct marks The means of the two distributions are identical because the mean is the balance point at which random errors cancel each other out

Total distribution The last columns in Table 8 1 and Figure 8 1, sum up all the ten frequency distributions The mean of course, remains 15 52 But the standard deviation of the total distribution of erroneous marks (7 187) does not have any easy interpretation and is quite possibly a figure without any practical meaning

7 *Interrelationships in Table 8 6(b)*

Some interesting conclusions can be drawn from an inspection of the interrelationships of the various columns in Table 8 6(b)

The means and standard deviations of the rank orders were calculated by pooling the data in Table 8 5 i e taking into account half ranks rather than the rounded ones in Table 3 4 It is interesting that these mean rank orders are almost perfectly correlated with the rank orders of the means (i e of the correct marks) Notice that, for example,

the extra large gap between the mean marks of A and B is reflected in an extra large gap between their mean rank orders However, the mean marks tend to "stretch out" more at the top end of the scale than do the mean rank orders This suggests that a scaling system based on ranks (as will be recommended in the chapter on Scaling) will quite fairly reflect the distribution of ability in the group, although it may lower the very highest marks slightly

The correct marks (means) and standard errors of marking (*s*'s) seem to be fairly highly correlated, the highest marks having the widest distribution of errors This means that there is wider disagreement among examiners as to what marks to award to the best students, than to the weakest

When we look at the rank orders, however, the situation is reversed The highest rank orders have the narrowest distributions This suggests that rank orders are much more reliable for the more able candidates than are marks

Finally, a comparison of means with medians suggests that there is relatively little skewness in the distributions, although there is some slight tendency for the high marks to be truncated at the upper end, and the low marks at the lower end This is as it should be for random errors, subject to the limitations near the ceiling and floor of the scale

8 Reliability

Any reader with training in educational measurement, or even some familiarity with psychological tests or objective examinations, will want to know, "What is the coefficient of reliability of this examination ? "

In the broadest sense, our entire discussion has been about reliability We have approached this in terms of variability in marks, in means, in pass percentages, in rank orders We have also seen (Chapter VII) that "*the* reliability ' is an ambiguous term, and that—when talking of traditional essay type examinations—we must think of at least three different types of reliability

We have studied extensively only *one* aspect of reliability—i e the reliability with which the examination is marked (It should be remembered that for objective examinations, there would be no variation in marking at all) Other factors which are known to lower reliability have not been studied variability in content (i e the traditional parallel forms situation), variability in performance (i e test-retest), the effects of allowing a choice of questions (In this examination, all candidates had selected the same five questions) Nevertheless, we will attempt to give some answer to the question about "the coefficient of reliability"

Examiner reliability Examiner reliability* is generally defined as the correlation between the marks that two examiners assign to the same set of answer books Since we have 90 examiners we have (90×89)/2 such pairs But the calculation of 4005 coefficients of correlation changing them to z, and averaging them is a rather large task! Therefore some simpler methods had to be found

Rank order correlations are easier and quicker to calculate than product moment r s, even though we lose some of the original irregularities of the marks distributions by translating them into ranks Furthermore, Guilford (1954 p 253) provides us with a modified rank-order coefficient of correlation [giving] an average intercorrela tion of all individuals' as follows

$$\bar{p} = 1 - \frac{k\,(4N+2)}{(k-1)\,(N-1)} + \frac{12\,\Sigma\,S^2}{k\,(k-1)\,(N)\,(N^2-1)} \qquad (8\ 1)$$

where $\bar{p}$ = average intercorrelation among individual judges
k = number of judges
N = number of stimuli
S = sum of the ranks for any stimulus

Applying this formula we obtained the following mean rank order correlations among the examiners

*An attractive idea presented itself from the data in Table 8 6 (and because of the graphic display in Table 3 1)—an idea so attract ve that it led the senior author down the wrong path for several months Luckily a brief talk with Harold Gull ksen clari fied the issue

It seems worth mentioning the idea here lest anyone else be misled The idea is that the standard error of marking represented by the frequency d stribut ons in Table 3 1 (and statistics in Table 8 6) is akin to errors of measurement in the statistical sense of the term (i e calculated from the usual formula) Thus it would seem that a comparison of the average of the standard deviations of the columns with the standard deviat on of the marks g ves us a class cal definition of the co- efficient of rel ability

The fallacy? The concept of rel ability is a *correlational* concept—and no correla tions are involved in the d stributions in Table 3 1 These d str butions reflect d ffe rences in standard of marking of various examiners They may of course also reflect lack of correlation between exam ners—but th s effect is thoroughly masked For example all of the top marks may belong to one examiner all of the med an marks to a second and all of the bottom marks to a third These three examiners would have reader rel ab l ties of 1 00—and yet produce a wide d stribut on of error in Table 3 1 The d stributions in Table 3 1 would show rel abil ty and measurement error in the techn cal sense only if all the marks had first been scaled to a common mean and standard deviat on

First order of marking (42 examiners) = 83
Second order of marking (48 examiners) = 85
Total (all 90 examiners) = 83 = examiner reliability*

A second approach was to correlate a number of actual pairs of examiners A 1% sample (40 examiners) was drawn by reference to Snedecor's table of random numbers (Snedecor, *Statistical Methods* (1940) pp 10—13) Reading across a line, chosen at random, beginning at a spot chosen at random, pairs of digits were read off to represent examiners *For example, the first series of digits read 803270267198* Thus we correlated examiner 80 with 32, 70 with 26, 71 with 98 The examiner numbers used were the original ones, which ran from 1 to 100 So the only restriction was that if the number of an examiner who had not returned his marks sheet was drawn, it was ignored and the next one drawn to replace him Thus the series 7631062905 would yield the pairs 76—6 and 29—5 as 31 had not returned his marks sheet

This is drawing "with replacement", and resulted in several examiners, and even one pair appearing twice If we had been sampling only the 90 examiners, we would have done it "without replacement", i e we would not have allowed any examiner to be included more than once The present method seemed justified because we were not really sampling 90 examiners—rather, we were attempting to draw a sample *representative of the total population of examiners, of which the* 90 were themselves only a sample

For simplicity, the correlation between each pair was again found by rank order methods, although this may obscure some of the results of irregular distributions that would have affected the product-moment Not all the figures have been checked (although the calculations are so simple as to make serious errors unlikely) However, all unusually low or unusually high values were double checked

The results are presented in Table 8 7 It is interesting to note that the median (86)** of this one per cent sample of rank order correlations is very close to the value of 83 found by formula 8 2

Remember, again, that 83 represents the reliability of marking *only* if the marks of all examiners are scaled to a common mean and

*Taylor *et al had computed Kendall s Coefficient* of Concordance for their data The result was 75 For comparative purposes we also computed it for our data obtaining a result of 836

**In the later computer analysis (mentioned in a previous footnote) of a stratified random sample of 49 answer books the median of 1176 intercorrelations of examiners was 88 Since the product moment *r* can be expected to differ from *rho* when examiners marks are not normally distributed, this seems to be a very close agreement indeed The median *r* for our sample of 40 pairs in Table 8 7 was 895 Taylor et al s mean *r* was 77

TABLE 87

Rank order correlations between 40 sample pairs of examiners, in the order in which they were drawn

Examiners	*rho*
80—32	86
70—26	98
71—98	89
19—18	92
43—42	95
90—81	74
34—92	96
68—95	87
44—11	90
54—96	86
20—16	93
80—92	.91
71—55	87
16—57	96
73—12	90
66—63	81
42—07	83
02—19	85
81—36	85
94—19	71
43—47	57
36—64	76
10—25	72
14—55	.79
48—02	.75
14—55	.79
97—48	88
22—88	80
13—66	84
34—59	.87
87—56	86
66—76	70
21—70	99
02—75	86
88—41	92
29—52	98
84—89	84
51—02	.91
72—30	93
56—46	84

Frequency distribution of rank order correlations among 40 pairs of examiners

rho	*f*	
99	1	
98	2	
97		
.96	2	
95	1	
94		
93	2	
92	2	
91	2	
90	2	
89	1	
88	1	
87	3	
86	4	
.85	2	
84	3	
83	1	
82		
81	1	
80	1	
79	2	
78		
77		
76	1	
75	1	Significant at 1% level
74	1	
73		
72	1	
71	1	
70	1	
69		
68		
67		
66		
65		
64		
.63		
62		
61		
.60		
.59		
.58		
57	1	Significant at 5% level

standard deviation Since Indian examinations are generally not so scaled (with the notable exception of Gauhati), this apparently high figure can be quite misleading It is worth repeating the example of an actual 83 correlation given in Chapter III, and adding some further statistics to it

Candidate	Marks by Examiner X	Y	Difference
A	7	23	16
B	5	15	10
C	8	27	19
D	20	34	14
E	11	29	18
F	6	26	20
G	2	16	14
H	7	17	10
I	10	22	12
J	5	13	8
Average difference			14 1

$r_{XY} = 83$
$\sigma_X = 4\ 66\ M_X = 8\ 1$
$\sigma_Y = 6.52\ M_Y = 22.2$

Standard errors of marking
of $X = 1\ 91$
of $Y = 2\ 67$
Mean of actual differences between marks = 14 1

Thus if X's marks were converted to Y's scale, the maximum difference between the two marks for any candidate is likely to be about 6*

The standard error of marking for our data as a whole is found, by formula (see Chapter VII), from the average standard deviation of 6 367 and the coefficient of examiner reliability of 83 This standard error of marking is 2 62. This would, of course, apply to scores on the same scale The contrast between this and the standard error of marking (unscaled) reported in Table 8 6 (3 66) shows a part of what is lost by not adopting proper scaling procedures The rest is shown by the large differences in means in Table 8 3

Content reliability We have already noted that examiner reliability is misleading because many assume that it means the same thing as objective test reliability But the examiner reliability of an objective test, of course, is 1 0 Thus, before comparing our reliability figure with that of an objective examination, we must find content and marks reliability (See definitions in Chapter VII)

Since we had selected ten answer books in which the candidates had all answered the same five questions, we were able to use Ebel's intraclass correlation (Ebel, 1965, Guilford, 1954) to find the content reliability of the examination For this calculation we chose the work

*The formula for the standard deviation of differences between marks is as follows

$$\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2 - 2\, r_{XY}\, \sigma_X\, \sigma_Y} \qquad (8.2)$$

of Examiner No 48, who went out of his way to write, on the back of his marks sheet "The marking has been done as objectively as possible. Each answer has been thoroughly read from each answer book before evaluation and its comparative worth has been determined as acurately as possible' The fact that his marks correlated 955 with the correct marks indicates that his care was worthwhile

The content reliability for this examiner is 85 As we will see later, content reliabilities can vary from examiner to examiner In the 'four thousand re examined' project, we will find that content reliabilities for this same history paper varied from 21 to 76 with a median of 48 In that project, however, different candidates had answered different questions, and that undoubtedly reduced the content reliability of the examination as a whole

Marks reliability As shown in Chapter VII, the reliability of the marks which the candidate is actually awarded is the product of the examiner and content reliabilities Thus we multiply 83 times 85, and obtain 7055 or 71 as the marks reliability This is the figure which we must compare with the reliability of an objective examination And as is well known, a reliability of 71 is considered quite inadequate for reliably differentiating between individuals *Thus the examination fails in the purpose for which it is supposed to exist*

The standard error of measurement (using the average standard deviation of marks 6 367) is 3 46 This means that *even if all papers are marked on the same scale*, there is one chance in three that any particular candidate's mark will be more than 7% below or above what his true mark should be For one candidate out of 20 the error will be 14% or more and for one out of every 100 (which is one thousand candidates in every lakh) the error will be 17% of marks or more And, because examiners differ so widely in the scales that they use, the actual errors are very much greater

As we have seen (Chapter VII) no exact solution of the problem of the *total* distribution of error has been found But we can at least state the *minimum* by adding to the SEM the average difference between the extreme examiner means and the correct mean This figure is 7 05, or 14 1% of marks (see Table 8 3) Thus for a substantial number of candidates, the actual error can be *at least* 21 1% of marks or more. The validity of this theoretical figure is shown by the fact that it is—as it should be—just slightly larger than the maximum difference between correct and obtained marks shown in Table 3 1

9 *How well did the original examiners do?*

It may be felt that the present group of 90 examiners—knowing that

they were participating in this experiment—may have marked the papers with extra care. The original examiners, marking under the usual conditions, may have been more careless. Is this true ?

Table 8.8 gives the data. The correct mark is, of course, the average mark of the 90 examiners. Against this standard, the original examiners* did about as well as the 90 re-examiners. The average of their ten marks was only slightly below the average of the correct marks (15.1 vs. 15.5). The widest error was about 4 marks. The average error of 1.9 marks is slightly lower than the average (2.5) of the 90 examiners. The rank order correlation between the correct marks and the marks awarded by these original examiners is +.89.

This cannot be compared directly with our average inter-examiner correlation of +.83. Rather, it is a correlation with the average of these examiners. Cureton (1965) provides a formula for the average rank order correlation of the examiners with the criterion. Translated into the symbols we have been using, and letting rho_{av} stand for the average rank order of the examiners with the criterion, Cureton's formula reads as follows :

$$rho_{av} = \frac{12 \Sigma yS}{k (N^2 - N)} - \frac{3 (N + 1)}{N - 1}. \qquad (8.3)$$

where the only new term is y, which is defined as the "criterion rank" of each answer book i.e. the rank order of its average mark.

Calculations for the 90 re-examiners give us a rho_{av} of + .90. The rho_{av} for the original examiners, +.89, is almost exactly the same.**

Taylor et al (1966 b) found that their original examiners correlated +.69 with the correct marks, while their mean r (tetrachoric) of their 19 examiners with the correct marks was .86. Neither the difference between Taylor et al's two coefficients, nor the difference between theirs and ours, is significant. This does not mean, of course, that it is incorrect to say that their original examiners correlated lower than our original examiners or their re-examiners—it merely means that no inferences can be drawn from these facts, about the general population of "original examiners" in history and English.

*There were six original examiners. One of them marked three of these answer books, two marked two each, and the remaining three marked one each. On the average, a sample taken from several examiners is likely to be closer to the mean than one taken from a single examiner. This is due to what statisticians call "regression towards the mean".

**In the computer analysis of a sample of 49 of the 90 examiners (mentioned in an earlier footnote), the product moment correlations with the correct marks ranged from +.801 to +.994, with a median of +.947. By definition, this is the "index of reliability" of the examiners Test theory tells us that the coefficient of reliability is the square of the index of reliability. $(.947)^2 = .89$, which, as we saw earlier, was the median inter-r of the 49 examiners.

TABLE 8 8

RELATIONSHIP BETWEEN THE MARKS AWARDED BY THE ORIGINAL BOARD EXAMINERS AND THE MEAN MARKS OF 90 EXPERIENCED EXAMINERS, FOR 10 HISTORY ANSWER BOOKS

Answer book	Correct mark	Orig nal examiner s mark	Difference
D	27 0	28	+1 0
E	23 1	22	−1 1
C	19 0	18	−1 0
I	15 5	12	−3.5
F	15 4	13	−2 4
A	14 5	17	+2 5
B	11 2	15	+3 8
H	10 9	10	−0 9
G	9 8	9	−0 8
J	8 8	7	−1.8
Means	15.5	15 1	−0 4
		Average d fference	1 9

Taylor et al (1966 b) found their original examiners had been slightly more lenient than the correct marks—ours were slightly stricter Ours awarded one less Second, and a slightly lower mean though their pass percentage (based on only 10 answer books) was higher, as was Taylor et al's But again, these differences are too small to draw any inferences beyond these particular samples

In general, then, we may state that there is no evidence that actual examination examiners, working under the usual conditions are any less careful than were the re-examiners in these two studies Furthermore, as we will state again later, the reliability of our Indian examiners seems to be slightly *higher* than the average in Western studies of this type Unfortunately, this still does not make this type of examination—either here or elsewhere—reliable enough for the purpose for which it is intended *In spite of the apparently careful work of the examiners, the examination marks do not differentiate between the true merits of individual students in any reliable manner*

10 Factor analysis (1) answer books

Through the kindness of Dr Ayres D'Costa, and the Ohio University computer laboratory, two exploratory factor analyses were carried out These were based on the randomly selected 49 examiners noted earlier An inspection of the means and standard deviations

obtained from this sample suggests that it was a good sample of the 90 experienced examiners.

The first factor analysis was based on the inter-correlations of the 10 answer books. These inter-correlations ranged from +.151 (between answer books D and J) to +.828 (between answer books F and G), with a median r=.568. It is interesting to note that the lowest correlation was between the highest and lowest ranking answer books. This suggests that they differed not only in merit but in other qualities also. But the highest correlation was not between answer books adjacent in merit, but between the candidates ranked 5th and 9th. These two seem to be viewed in the same order of merit by all 49 examiners.

The first unrotated factor accounted for 59.01% of the variance. The second factor accounted for 11.80% of the variance. The third factor accounted for 9.85% of the variance. The fourth and fifth factors accounted for 5.45% and 4.16% respectively, and the last four factors 2.58%, 2.30%, 2.23%, and 1.47% respectively.

Seven of the answer books had negative loadings on one or more factors.

No attempt has been made to interpret these factors. This would be an entire study in itself, involving subjective ratings of the answer books by experts on several hypothetical variables (e.g. order of merit, handwriting, organization, style, grammar, spelling, etc., etc.) and possibly a re-analysis including scores for these variables.

What we can conclude is this : The process of marking is obviously a complex one, and a mark *does not mean the same thing for different candidates*. The major factor is presumably "knowledge of history", but a surprisingly large 41% of the variance is due to other factors. What leads to the raising of the mark on one answer book may be irrelevant in a second, and may even lead to the lowering of the mark in a third (hence the negative loadings).

We hesitated to report this, because such a partial report (without the trappings of correlation and factor matrices, factor leadings, etc.) may be frustrating to the expert. (The figures are available to anyone who asks for them.) But we felt that partial information is better than none at all. The study definitely shows the value of further, more detailed research along these lines.

11. Factor analysis : (2) examiners

Correlating 49 examiners on only 10 variables is, of course, somewhat absurd. As such, the results should be considered to be *of exploratory value only*. This is the kind of exploration which could not possibly be justified with only a calculating machine available, but which the speed of the computer does make possible.

The first factor accounted for a surprisingly large 88 14% of the variance Most examiners had a loading of over 9 on this factor, and the lowest loading of any of the 49 was 81 Thus this is a general factor, showing a reasonably high degree of overall agreement among the examiners This is the same agreement which is shown by our mean rank order correlation of + 83

This is not necessarily contradictory to the findings in the previous section It indicates that whatever the differences between answer books they perhaps balance each other out so that the overall order of merit remains approximately the same from examiner to examiner, for the batch of answer books as a whole A low correlation between two answer books means that they do not maintain a constant relationship to each other—but an unstable relationship between two out of ten does not have much effect on the correlation between two examiners based on the ten It is also probable that the answer book factors were correlated (Rotation was not carried out)

This first factor is the only one that we can interpret with any certainty The correlation matrix was based on 50 variables the marks of 49 examiners plus the "correct mark' (the mean of the original 90 examiners but rounded to a whole number) In the first factor, the loading of the correct mark was + 999 Thus the first factor is quite probably "knowledge of history , which is the only single thing that all examiners would agree that the examination is supposed to measure

The remaining 12% of the variance is accounted for by eight factors only four of which account for more than 1 5% each Each of these is bi polar More important, the loading of the correct mark on none of them is above ± 02 Thus there are certain qualities which attract certain examiners and not others—and the individual tastes of examiners tend to cancel each other out so that they have no effect on the average of their marks For example the mark on an answer book may be raised by one examiner because of evidence that the candidate has carefully revised it before handing it in—while another may lower the mark because crossing out and re writing destroys ' neatness Of course we do not know what the bi polar factors are This would take very extensive research Probably we would have to start with hypotheses built on the factor analysis of the answer books Then we would need a questionnaire asking each examiner how he thinks he weights (or intends to weight) various factors, such as handwriting neatness number of pages grammar, knowledge of facts organization, creativity and imagination etc Correlating the questionnaire results with the factor analysis might then give us some insight into the factorial ways in which examiners differ

In conclusion This factor analysis is presented just to show that

examiners do differ. How they differ is a major research subject in itself.

12. Position effects

The experimental design—having half the examiners mark the answer books in one order, and the other half mark them in the reverse order—was intended to bring out position effects. None were found that were statistically significant. Table 8.6(a) makes it evident that there was no significant difference between the first-order and second-order marking of each answer book. When we pooled data on three answer books—i.e. A, B, and C marked first and marked last—the difference was still not significant. Nor was there any consistent tendency for the mark of a particular answer book to be raised or lowered by the "contrast effect" with the answer book which immediately preceded it.

It is possible, of course, that the investigators did not apply the right hypotheses. For those with promising hypotheses, the data are available in the tables. However, note that Taylor et al (1966 b) found position effects showing up only after about 25 of their hundred answer scripts had been marked.

9

"Four Thousand Re-examined" —Detailed Reports

THE NEXT five chapters will cover the research on four thousand answer books, the work of 44 examiners, 11 in each of four subjects The highlights of this study have been given in Chapter IV, and will not be repeated The data and experimental design were discussed in Chapter VI

Chapter X will compare the four examinations history, Hindi, biology, and mathematics (geometry)

Chapters XI to XIV discuss the four subjects in detail It is assumed that many readers will be interested in only one particular subject Therefore, at the risk of seeming repetitious, each chapter is written to be complete in itself However, cross references to interesting points in other chapters are occasionally given

After reading Chapter X (comparing the four examinations), the reader may thus read just the one subject most interesting to him, and then skip to the Recommendations (beginning with Chapter XV)

A note on the t-test

A special note is relevant here on the interpretation of the "significance" stars in the subsequent tables The t-test for the significance of a difference between means is supposed not to be valid if the variances are not homogeneous Unfortunately, in quite a few cases the difference in variances is significant However, we are dealing with correlated means, and there seems to be no easy way to handle this problem Since

our *t*'s have been calculated from the distribution of actual differences (rather than from the formula involving *r*), we take them to be reasonable estimates. A more exact approach is likely to show a larger number of significant differences, rather than a small number.

10

Comparing the Four Examinations

THIS CHAPTER will give the over all comparisons of the four examinations To avoid repetition however, material which has already been adequately covered in Chapter IV will not be referred to again Only two of the topics discussed there appear here in a more detailed technical form

Table 10 1 tells the main story This is a summary of the tables given in each chapter Each chapter gives the data for each pair of examiners 10 pairs in which the same examiner re-marked his own answer books, and 10 pairs in which the answer books were re marked by an examiner other than the original one Thus there were 20 pairs for each of four subjects—80 pairs in all

Let us look at Table 10 1 in detail

The first three columns present results for the History examination In the first column are the averages for the 10 sets of 50 books each re marked by the same examiner In the second column are the averages of the 10 sets of 50 books each which were marked originally by one examiner, but re-marked by a different examiner The third column gives the averages for both sets of examiners

We not only have to distinguish between two types of second examiners, we also have to distinguish between the first and second marking O means the original or first marking Remember, *this was the regular marking of the answer books*, done under annual examination conditions The re marking experiment was not even conceived at the time that this original marking was done R refers to the re marking the second marking

the marking done for this experiment The examiners were asked to do their re-marking as much as possible under usual annual examination conditions To do it *exactly* the same way, knowing that this was an experiment in re-marking, obviously could not have been possible Still, so far as our evidence goes, there seems to be no reason to believe that examiners were extra careful, or otherwise "non-normal' the second time

It should be noted that *all marks are out of 50* The Hindi exam had only 34 marks, but all calculations were multiplied by 50/34 to make them comparable to the others

1 Mean marks

Now look at the two lines marked Mean marks O and R These are the averages (means) of 20 means, each based on 50 answer books The two Arts subjects have roughly the same average marks about 17 5, or 35% The two science subjects are higher and approximately equal 21 marks out of 50, or 42% These are substantial differences, which may or may not be related to any real differences in ability (See discussion of sampling in Chapter VI) We know of no studies directly comparing marks and ability levels of Arts *vs* Sciences students However, the senior author performed some comparisons in some tables given by Liddle (1965) He found that almost every combination of Verbal and Quantitative ability predicted higher Science than Arts marks *Since* Liddle's thesis was based on an excellent sampling procedure, this strongly suggests that students of *equal* ability will have *higher* marks if they offer Science than if they offer Arts Thus the differences noted here are quite probably due at least partly to the use of different scales, and not to any real differences in ability between Arts and Science students

The line labelled *Absolute differences* gives the average difference between the 10 (or 20) averages *without regard to sign* The following artificial example will make this clear

		Same	D ff	Total
Mean marks	O	24 00	20 00	22 00
	R	20 00	24 00	22 00
(D fferences)		—4 00	+4 00	nil
Absolute differences		4 00	4 00	4 00

When each of a given pair of examiners marks an answer book independently, there is no *logical* reason to call one "first' and the other "second" Logically, we just have two independent markings And if

we decided purely at random which was 'first" and which "second," their average differences would end up around zero Therefore, what we are interested in is the difference regardless of which is called 'first" and which "second" This is what "absolute differences" means

Let us go back a moment We pointed out that only one of the two markings was under "real" conditions the other was under, shall we say, "simulated real" conditions And we can see from our Table 10 1 that there was *some* consistent difference between the two markings Some of the difference in means is due to reduction of the 'J-effect , i e of the piling up of marks just at the pass mark of 33% As can be seen from Table 4 5, many of these marks were re-distributed downward, thereby lowering the mean However, on the average the re-examiners were also a little bit harder on the candidates than the original examiners were Perhaps they were being more cautious ? Or perhaps knowing that these marks would not actually be awarded the candidates, they felt they could afford to be less lenient ? We do not know the exact reason However, we *can* see that the differences are small, the largest (biology) being only 0 83 marks out of 50* Because of the large N's (1,000) three of these four differences are *statistically* significant However, it is unlikely that they are *psychologically* (or educationally) significant (Admittedly this point is arguable, when a single mark in error can make the difference of pass or failure to literally thousands of candidates However, these differences are minor compared to the other sources of error in this examination system)

We now return to our 'Absolute differences' For the Totals, these are largest for history (2 13) and nearly equal for the other three subjects There is a rather remarkable degree of agreement in "standard of marking" between pairs of examiners, *on the average* But the city of Allahabad has not been saved from flooding just because the *average* rainfall is only one inch per week Similarly, reference to Table 11 2, 12 2, 13 2 and 14 2 will show that there is a wide variation among examiners One pair of history examiners differed by an average of 5 marks (10%) on 50 answer books, while one biology examiner disagreed with himself by as many as 4 64 marks (over 9%)

Look at the Absolute differences in the 'Same' and 'Different" columns In biology (partly because of the above-mentioned examiner) there was more agreement between different examiners (1 19) than between the two markings by the same examiner (1 39) In the other three subjects examiners agreed more with themselves than with each other as to the average merit of the answer books.

Remember, of course, that these are average marks on 50 answer

* Because they are unimportant these differences are not given in the table If the reader desires he can easily subtract one mean from the other to obtain them.

books. Disagreements on individual marks ran as high as 20, as was shown in Table 4.1.

2. Standard deviations

The average standard deviations tell roughly the same story. Perhaps to make up for the lowered means, and to avoid penalizing the best candidates, three of the four standard deviations were increased somewhat. Some of this increase may also be due to the same factor as for the mean, i.e. the reduction of the J-effect. (In Hindi, which did not increase, the J was at the mean.) However, since standard deviations depend on the square of the deviation from the mean, a shift of a point very close to the mean would not have very much effect. And the fact that this is not the only cause of the increased standard deviations is indicated by the fact that the percentages in upper divisions, and the percentages failed, also changed from Original to Re-marking.

But the difference between "Same" and "Different" examiners is much more striking here than for the mean marks. The "Different" examiners had from two to four times as much difference between them as the "Same" examiners did. This suggests that, though there is a fair degree of agreement in the average standard, there is far less agreement among examiners in the awarding of very high and very low marks. This may be because only "pass percentages" are published, and given much publicity. Thus the examiner has a fairly good guide as to how many should pass—but not how many should get Second or First Divisions—or, conversely, how low the lowest should be marked. If this hypothesis is true, *better instructions to examiners* are needed to cure this defect. (Every American teacher we have ever met knows what "marking on the curve" means, and so do most American high school students. Thus most American marks tend to be distributed normally, with the proportions in each of the various "divisions" being approximately equal for all subjects. It is tragic that this very simplest of statistics is apparently unknown, or at least unpracticed, in India.)

The standard deviations of the four subjects tell an interesting story—one that has already been mentioned in the Summary section. The standard deviation of Mathematics is twice that of any of the other three subjects. *Mathematics answer books are marked on a different scale from other subjects. It is as illogical to compare the marks of mathematics students with those of other students, as it is to expect a 39-inch table and a 39-centimeter table to be of equal length.* This has been pointed out repeatedly in India (Mahalanobis 1934, Bose 1955, Harper 1963, and others). Is it not time to apply a more rational scaling system ?

Particularly devastating is the fact that it is the *standard deviation,*

not the mean, which controls the actual weight a particular mark receives in the total aggregate marks (See any textbook on psychological or educational statistics, also Harper 1963) This gives the good mathematics student a tremendous and basically unfair advantage in competing for "firsts" and other prizes, and the poor biology or Arts student a rather large disadvantage If we place together a 40 inch table and a 40-centimeter table, is the total length 80 ? *Examination marks with such widely different standard deviations should NEVER be added* They should first be scaled to a common scale

3 Actual vs intended distributions

It is interesting to compare the standard deviations in Table 10 1 with those in Table 4 5 Table 10 1 reflects rather large variations from examiner to examiner in "standard ' But we may take the average of the standard deviations of these examiners as their consensus as to how marks *should* be distributed Let us compare these *intended* distributions with the *actual* distributions, as shown in Table 4 5

	History		Hindi		Biology		Geometry	
	O	R	O	R	O	R	O	R
Intended σ	4 14	4 88	4 66	4 43	4 81	5 14	10.34	10 81
Actual σ	4 46	5 62	5 01	5 19	5 12	5 37	13 42	13 33
Increase	32	74	35	76	31	23	3 08	2.52

In every case, the actual standard deviation for the total group is larger than the examiners intended This, of course, is the result of tabulating together the unscaled marks of examiners with different means and standard deviations, i e examiners who are actually using several different scales * Thus the fact that the standard deviation of all the marks in a subject is larger than what any single examiner intends it to be is just one more defect of the examination system, due to lack of scaling This has the effect of scrambling the positions of candidates in unpredicted ways To put it in technical terms the percentile rank of the candidate in the total group is often very different from what his particular examiner intended it to be

Furthermore, as we have already noted it is the standard deviation (distribution) of marks in any one subject which determines that subject s

* The mechanism involved can most easily be demonstrated by considering ranges (rather than standard deviations) Suppose Examiner A has awarded a range of marks from 10-50 with a mean of 30 Suppose Examiner B has awarded to a group of similar ability, marks ranging from 30-70 with a mean of 50 When we combine these two the new range is 10-70 which is one and a half times as large as either examiner considered relevant to these groups

relative weight in the total aggregate marks It is this "enlarged by error" standard deviation which helps determine the position of any individual student in the aggregate total—not the intentions of any single examiner, or even the average intentions of all the examiners

Notice also in the above table, that the largest increase in the distribution (σ) is in geometry, which is supposed to have the least subjective marking The actual standard deviation of geometry marks is nearly three times that of any other subject—thus giving geometry nearly three times the weight in the aggregate marks, that any other subject carries Is it really the intention of those who control the examination system that a student's ability in mathematics should be three times as important as his ability in any other subject, in determining his final rank ?

4 Note on use of medians for correlation coefficients

All of the above figures have been arithmetic means For the correlation coefficients which follow, medians have been used Partly this is because averaging requires translating into z first, which is very time consuming when 360 coefficients are involved Partly it is because coefficients based on 50 cases are not highly reliable, and thus "chance" variations can be quite large Use of the median ensures that any "chance" variations at the extremes will not unduly influence the average

5 Correlation of no of pages vs marks

Does the number of pages written influence the number of marks awarded ? The evidence that, in general, it does not is discussed more fully in a later section (see Table 10 4) It can be seen here that the medians are quite low—no higher than they should be on the reasonable hypothesis that, after all, the worst student *knows* less than the best, and can be *expected* to write fewer pages

Note that there seems no particular logical reason to expect the correlations for "Original" marking to be different from those for "Remarking" And the medians obtained support this

6 Mean of absolute differences

This is put way down here to try to discourage anyone from confusing it with the "Absolute differences of means", given above

The absolute differences in the section on means were the absolute differences between means or averages

The absolute differences here are the absolute differences *between individual marks* In other words, we have added up the difference between the first and second marking of each individual answer book,

and averaged these They are "absolute differences" because they are without regard to sign As we have pointed out, there is really no logical reason to consider one marking "first" and the other "second', so a difference is a difference, regardless of its direction The following example will make this calculation clear

Candidate	First Marking	Second Marking	Difference
a	43	39	—4
b	37	35	—2
c	40	42	+2
d	25	29	+4
Total	145	145	12
Mean	36.25	36 25	3

Note that in this example there is no difference between the means (36 25), but the mean of absolute differences is 3 This shows the difference between the two concepts

Look at the four "total" columns The largest difference between examiners is in history (an average of 3 05 marks out of 50, or 6%) and the smallest in mathematics (2 58, or 5%) Notice, however, that there is really very little difference in the accuracy of history and mathematics marks—in spite of the popular superstition that mathematics marks are far more accurate than history marks Of course, if history and mathematics were marked on the *same* scale, then our other evidence shows that the mean of absolute differences would be *much* less for mathematics than for history But as the marks stand, mathematics cannot boast any higher stability for its marks than history

Now look at the "Same" and "Different" columns In history and in mathematics, examiners re marking their own answer books do substantially better (2% of marks) than when they are re marking another's answer books For Hindi and biology, examiners do not seem to be agreeing much better with themselves than they do with each other, so far as this measure is concerned

Incidentally, do not take comfort in the fact that these differences look rather small Remember, these are just averages The average temperature of Delhi may be 75°—but this won't save you from the possibility of heat stroke in June Table 4 1 shows that some of the differences contributing to these averages are as large as 40 per cent of marks

7. Examiner reliability

Examiner reliability is the product-moment correlation between

two examiners marking the same set of answer books. It is the measure of agreement between examiners. It is also known as "reader reliability".

We found a range of examiner reliabilities (both self-consistency and inter-examiner consistency), of course. Perhaps the surprising thing is that the range was not larger. Here are the figures for the range of correlations, when the same and when a different examiner re-marked each batch of 50 answer books:

	Same	Different
History	.72 to 93	.49 to .85
Hindi	61 to .94	.66 to .88
Biology	.75 to .91	.67 to 88
Mathematics	81 to .99	92 to .97

We noted in the last section that Hindi and biology examiners do not agree much better with themselves than they agree with each other—while for history and mathematics examiners there is a substantial difference. Exactly the same picture is shown by the examiner reliability coefficients. But to see this, we must translate the coefficients into Fisher's z, because a difference of .03 between .95 and .98 is actually 50% larger than a difference of .11 between .74 and .85.

Subject	History		Hindi		Biology		Mathematics	
r or *z*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*
Same	85	1 256	.81	1.127	.83	1.188	.98	2 298
Different	.74	950	.78	1 045	.80	1 099	.95	1 832
Difference		.306		.082		.089		.466

Thus we see that mathematicians agree more highly with themselves than with each other in marking papers, and so do historians.* But for Hindi and biology the difference between "self" and "other" agreements (i e. between the self-consistency and inter-examiner consistency types of reliability) is negligible.

Why these differences? The real reasons could be known only with further research; however, it may be of some value to speculate. We can rule out memory as a factor, because of (a) the substantial number of answer books and (b) the time interval of nearly nine months A more reasonable hypothesis is in "styles of marking". The data seem to suggest (see next section, also chapter VIII) that different examiners are looking for different things, and that these differences are consistent

* The statistical significance of these two differences is beyond the 1% level of confidence. However the *educational* significance of the differences is probably negligible The number of students whose class or divisions would be different for Same and for Other re-marking is small

traits for examiners For example, one historian may be more impressed with accurate dates another may overlook an error in date if the candidate is approximately correct and shows a real understanding of the era and events involved Mathematicians may give more or less credit for certain aspects of partially correct answers The data suggest that perhaps there is wider agreement in Hindi and biology about what should and should not be credited than in the other two subjects Better instructions to examiners can reduce these differences

The "Total" columns tell a different sort of story Though mathematicians may agree better with themselves than with others they do agree with each other (95) much better than other subject-examiners agree even with themselves (85 81, 83) Thus examiner reliability for history, Hindi, and biology seems to be about equal—but for mathematics it is definitely better Of course even for mathematics it is *far* below the "examiner reliability" of objective examinations, as the latter is 1 00 In terms of z, the difference between reliabilities of 96 and 9965 (which is the closest to 1 00 that the table comes) is *larger* than the difference between reliabilities of 0 and 80

These examiner reliabilities are in line with those found in Western studies—except, of course, for those studies which are based on special and intensive training of examiners, and greatly improved marking systems There is a general belief that Indian examiners spend less time per paper than in the West (which may be true) and that they are therefore more careless and less reliable The latter, apparently, is not true *Indian examiners are apparently no less reliable than Western examiners But traditional examinations in themselves are unreliable*

Thus we cannot take much comfort in the reader reliabilities which we have found for history, Hindi and biology Gulliksen (1950 p 212) remarks "In general we should strive for a reader reliability over 90 if it is possible to achieve this level It would seem that a reader reliability of less than 80 is so low as to necessitate further discussion and alteration in methods of reading" Only one-quarter of our inter-examiner consistency figures were above 90 (and those mostly in mathematics) while nearly one half (19 out of 40) were below the 80 minimum

8 *Content reliability*

Not only are these examiner reliabilities inadequate—they also can be misleading As we noted in the last chapter, coefficients of correlation take no account of differences in mean and standard deviation Thus one examiner could fail an entire batch of candidates and the other pass all of them with high marks—yet their examiner reliability could be as high as 1 00, if they both put the answer books in the same order of merit

This factor could, of course, be taken care of by scaling the examination marks—as is recommended several times in this book A more serious problem, therefore, is the fact that all too often examiner reliabilities of essay-type examinations are talked about as though they were the same as objective-examination reliabilities It has often been shown that examiner reliability can be substantially improved by more carefully defined marking standards, better techniques, better training of examiners, etc (See Basumallik, 1959, for an excellent review) The assumption often seems to be—perhaps unconsciously—that "If we can raise examiner reliability to 90, the essay examination will be as reliable as an objective examination" Nothing could be further from the truth Though an essay examination cannot be reliable without high examiner reliability, *high examiner reliability in itself does not make an essay examination reliable* Remember that objective-type examinations have an examiner reliability of 1 00 Even an utterly unreliable objective-type examination (we once actually saw one whose odd-even reliability was 0!) still has an examiner reliability of 1 00

The coefficient for an essay examination, which must be compared with objective examination reliability, is what Gulliksen (1950) calls "content reliability". The reliability coefficient for objective examinations is based on the contents of the examination, not on agreement between examiners The same must be required of essay-type examinations

The various ways of obtaining the coefficient of content reliability have already been discussed in Chapter VII An abstract of Harper's adaptation of Ebel's intra-class correlation is given in the Appendix It was also noted, in Chapter VII, that this method may produce spuriously high content reliabilities, due to the "halo effect"— an effect which is not discouraged by the marking methods generally prevalent in India (i e reading through each entire answer book, rather than marking Question 1 in all answer books, then Question 2 *in all, etc*) Thus it seems reasonable to consider our obtained content reliabilities as upper limits, under current conditions

The fact that a wide range of reliabilities was obtained for each examination, though the question paper was exactly the same in every case, indicates that what is being marked, how much relative weight is given to various elements, etc , must differ from examiner to examiner. The ranges and median content reliabilities were as follows:

Subject	Range	Median
History	21 to 76	48
Hindi	45 to .77	70
Biology	41 to 68	63
Mathematics	71 to 90	77

It is instructive to compare these history reliabilities with the one reported *for the same examination* in the ninety marking ten experiment A choice of five questions out of ten was allowed the candidates but for that experiment answer books were selected in which the same five questions had been answered The content reliability for a batch of answer books in which all candidates had answered the same five questions was 85 (Chapter VIII section 8) When different sets of five out of ten were answered the maximum content reliability drops to 76 and the median to 48 This supports the recommendation of all experts (DFPSE 1963) that all questions should be compulsory

We have not found any other content reliabilities with which to compare our history Hindi and biology results Gayen (1961) found the reliability of a similar mathematics examination to be 63 Considering the difference in techniques these coefficients are reasonably close Note that for mathematics content reliability is considerably lower than *examiner reliability—so high reader reliability is not much use* when the examination method itself is defective This points up the fact that the main advantage of so-called objective tests is *not* their objectivity *in marking* (That exists *in Mathematics also*) The *major strength of* objective examinations is in their *much more adequate sampling of the candidate's knowledge and understanding of his subject* It is here that the traditional type of mathematics examination *with its very limited* number of questions falls down completely The elimination of choice and options a substantial increase in the number of questions in mathematics and the wider use of short answer (*or an improved form of short notes*) *questions in the other subjects* could raise these content reliabilities significantly

We have no other reported reliabilities with which to compare our results It is interesting that so crucial a question as the reliability of examinations where a choice of questions is allowed has apparently never been tackled before Even Gayen (personal communication 1966) side stepped the choice problem by simply lumping together all options into a single score for the group to which they belonged The present writer considered this question so crucial that he held up this entire research project for several months searching for an answer

One last comment on content reliability *No objective examination with reliabilities as low as these would be considered acceptable for determining examination marks*

9 *Marks reliability*

If we are to compare essay with objective-type examinations it is the reliability of the actual marks themselves which is *most* important

For objective type examinations, this "marks reliability" (also called 'score reliability') is identical with "content reliability", which is why we have only a single term or name, "reliability" But for traditional examinations, the content reliability is reduced or attenuated by the fact that two examiners will not give the same marks Thus, marks reliability is the *product* of examiner reliability and content reliability (This formula is derived from Gulliksen, 1950, pp 211-214 See Chapter VII) Thus, if examiner reliability is 90 and content reliability is 80, marks reliability will be (90) (80)= 72

The reason marks reliability is given in the "Different" column in Table 10 1 is this What we are interested in is the reliability when two different examiners are marking the test—since very few students have the same examiner, and most of them have different examiners Therefore, for each pair of examiners, we multiplied the "Diff' examiner reliability (i e the inter examiner consistency) with the content reliability to obtain marks reliability The figures given in Table 10 1 are the medians of these 40 calculations* Figure 4 5 plots the individual coefficients from which these medians were derived

None of these reliabilities would be considered adequate for an objective examination *Not one of these examinations is adequately reliable for the use that is made of the results*

10 Comparison with other studies

A question has been raised about the relationship between the "marks reliabilities" found in this study, and the intercorrelations of examinations in other studies For example, the Maharashtra Board reports correlations of 53 and 52 between the Preliminary Examination (held by the schools) and the S S C for History Does not such a result call into question our reported reliabilities of 28 to 43 for History?

It may be mentioned that most such correlations reported in the literature are quite low Still, the problem of these occasional higher ones must be dealt with Similarly, some studies have shown predictive validities for objective tests predicting essay exam marks, to be higher than the essay-exam reliabilities that we have reported Can validity be higher than reliability?

The last problem is the easiest to deal with When two examinations are correlated, the upper limit of the coefficient is equal to the square root of the product of the two reliabilities (Cf the formula for the correction for attenuation) Thus, as the reliability of one of these variables

*Table D in the Appendix gives all the details which entered into the calculations Examiner, Content and Marks Reliability

(say, the objective examination) approaches unity, the validity coefficient approaches the *square root* of the lower reliability. Thus an examination whose reliability is only .28 can have a correlation approaching .53 with a highly reliable examination.

Secondly, most of the published studies deal with subject marks, while ours deals with marks on only a single paper. The reliability of a mark which includes two papers will be higher than that of either paper, (except in the unlikely event that the correlation between the two papers is zero).

Thirdly, results are strictly comparable only when the work of a single examiner (or of a single pair of examiners) is examined. The effect of range on correlation coefficients is well known. On the average, the standard deviation of the marks of any single examiner will be less than the standard deviation for the examination as a whole—because the latter is made up of marks from many examiners, whose standards differ. Thus the pooling of results of two or more examiners may artificially either raise or lower the total correlation An example of the former case is given in Figure 10.1. Note that Examiners A and B are *stiff markers*

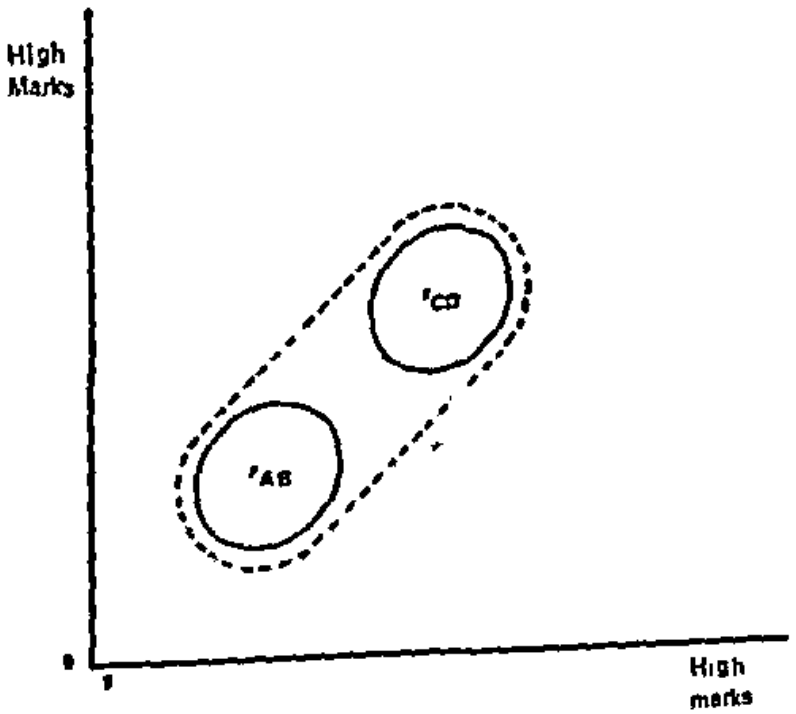


Fig 10.1

for their batch of papers (or else they have drawn a poor sample of candidates), while Examiners C and D are lenient markers for their batch (or else have drawn only good candidates). The correlation between each pair of examiners is low—but when the data are pooled (dotted line) the total correlation for the scatter chart becomes quite high. Similarly,

if the results of two examiners, both with low reliability, are pooled, it is quite possible that the correlation with the criterion may be considerably higher than the two reliabilities.

In such cases, the standard errors of measurement and standard errors of estimate would probably be a much more stable basis for comparison than the correlation coefficients, which are subject to so many irrelevant influences.

Finally, let it again be pointed out that Harper's reliability formula for essay examinations is new. It has not yet been subjected to "the test of time", and in fact has been subjected by only one researcher to an empirical confirmation of its validity. It is not impossible that the formula may be less accurate than logic, at present, leads us to believe.

11. Standard error of measurement

Neither the student nor the teacher is interested in reliability coefficients. What the student wants to know is, "If I took the examination again, and it was marked by a different examiner, would I get the same mark?" This is what the last line in Table 10.1 attempts to answer.

Reliability coefficients are affected by some factors that have nothing to do with reliability. This is because they are correlation coefficients. Particularly, they are affected by the range of talent in the group taking the test. It is quite possible to construct a test which will have a reliability of .95 when administered to children ranging in age from two to sixteen—and yet have a reliability of .40 or less for any single age group. The standard error of measurement, on the other hand, will remain relatively constant regardless of the range of talent examined. Thus, the standard error of measurement is generally *a far more important and realistic indicator of the true reliability of a test in actual practice*, than is the coefficient of reliability. (Harper 1959, Varma 1962, and standard statistical texts.)

The standard error of measurement (SEM) is the standard deviation of error scores around the "true score" for any test. If we administered an infinite number of essay examinations to a single student, and had them marked by an infinite number of examiners, we might consider the average of all these marks the "true mark" for that student. And the standard deviation of the infinite number of marks, around this true mark, would be the SEM of the examination.

More practically, the SEM enables us to make statements like this: "In mathematics (see Table 10.1) there are two chances out of three that a candidate's mark will be 5 or less points away from his true mark. There is one chance out of six that it will be more than 5 marks too high; and one chance out of six that it will be more than 5 marks too low."

Again remember that these are average or mean SEM s As can be seen from Figure 4 6 the error for some examiners is much larger

As has already been pointed out (Chapter IV) mathematics has the largest SEM (in spite of having the highest reliability) because it uses a different scale from other examinations one with a much wider standard deviation If mathematics were scaled to the *same* scale its SEM would be the smallest of the four This shows clearly why the SEM is *a better indication* of the reliability of actual marks than is the coefficient of reliability

Another fact must be pointed out These SEM s are actually minimum estimates They are based on the mathematical assumption that both examiners of each batch of papers used the *same* mean and standard deviation This we know is not true—and the differences in M and σ would inevitably make the standard deviation of error much larger than these SEM s in actual practice

12 Scatter charts for Same and Different re examiner

Table 10.2 shows for each of the four subjects how examiners differ from themselves and how they differ from each other when answer books are marked again In each subject 500 answer books were re examined by the original examiner and 500 were re examined by a second examiner (These data were pooled in Table 4 3)

The interested reader may want to make the same type of detailed analysis of Table 10 2 as was done of Table 4 3 (see Chapter IV section 4) But perhaps the most striking thing about Table 10 2 is just the demonstration of the fact that experienced examiners re-examining answer books several months later disagree with their own first markings almost as much as they disagree with each other

Also of importance is the number who change from pass to fail and from fail to pass The reader should count these for himself to bring home to himself the extent of human tragedy which can be caused by the unreliability of the traditional type of examination

13 Six examiners compared on equivalent answer books

There is a possibility that some readers may think that some of the results shown thus far may be partly due to the fact that each examiner receives answer books from a different batch of candidates These batches we know may differ from one another in ability both average ability and range To counter this possible source of error—and also to sum up the study more precisely—we used the following procedure for a special part of the total study We drew 300 answer books from the work of one specially experienced examiner (Examiner X) in each subject These 300 answer books were divided into six sets of 50 each which were as

nearly identical as possible in mean and standard deviation. Each of these six sets was assigned to a different examiner (Examiners E, F, G, H, I, J). As with all examiners in this project, they received the same instructions as were sent to the original examiner. Table 10.3 and Figure 4.2 give the results of this special part of the experiment.

Figure 4.2 has already been discussed (Chapter IV, section 6). The data in Figure 4.2 and Table 10.3 are the same, but the latter emphasizes means and differences in marks while the former concentrated on differences in the ranges of marks awarded equivalent answer books, by different examiners. For a more stable figure than the actual range itself, the standard deviation was used. Thus each bar actually included a little less than half the range.* Figure 4.2 was arranged in order of standard deviation, while Table 10.3 is arranged in order of mean.

In Table 10.3 you will find the result for Examiner X (the specially experienced examiner who originally marked all 300 answer books) listed first in each section. You will then find the results of the six re-examiners, listed in order of average marks. Means and standard deviations are given. The third column shows the average difference between the two marks awarded, one by Examiner X and the other by one of the six re-examiners.† And the last column shows the single largest difference, in each case.

The reader should not take much comfort in the fact that the "average difference" (mean of differences) is so much smaller than the "largest difference". Averages are misleading. We must not forget the man who was being tortured by the Gestapo. They had his head in an oven and his feet on a cake of ice. When asked how he felt, he replied, "On the *average*, I feel fine." One psychologist, in fact, strongly advised us that we should not report the average differences at all. It is the largest differences which, after all, are far more important. Remember that *each* "largest difference" is the change in the marks of one answer book out of 50. This large difference represents, then, the *minimum* difference in marks for *two per cent* of all answer books for these two examiners. Considering that an examiner may mark from 400 to 500 answer books,

* Each bar included $\pm 1\ \sigma$ from the mean For N=50, range/σ averages 4.5. (Guilford, 1956, page 43). Therefore each bar was about 44% of the range, though it included 68% of the cases

† Note that this "mean of the differences" is not the same as the difference between means. As was explained in more detail earlier (see "Mean of absolute differences"), this was obtained by adding up all differences without regard to sign. For example, a difference of 3 was counted as "+3" regardless of whether the second mark was higher or lower than the first. But the differences where the second mark is higher and those where the second is lower tend to cancel each other out to some extent, making the "difference between the means" lower than the "mean of the differences".

this error of 16 marks represents the minimum†† error for *eight to ten of his candidates* If *your* son has failed instead of being awarded a Second Class (which is what an error of even 7 marks out of 50 can mean) it will not make you any happier to be assured that the *average* error is only 2 or 3 marks!

The last line of each section in Table 10 3 gives the range from highest to lowest value in each column, i e the difference between the highest and lowest means and SD's These figures are good indicators of the amount of subjective judgement that goes into fixing the "standard ' in any subject—or, rather, the extent to which examiners' subjective judgements of ' standard" differ

History shows the largest variation in averages (Table 10 3 gives it as 7 44, and you can see how they jump around in Figure 4 2) with Hindi and biology next However, biology shows by far the largest variation in standard deviations (3 577) Comparing these with Hindi and geometry, suggests that the variable standards in history and biology are due to the greater amount of judgement involved in marking As every lawyer knows, human judgement is highly fallible Any newspaper reader knows how often the carefully considered judgement of one court is overthrown by a higher court Many of the techniques of modern examination methods are attempts to reduce the effect of this judgement factor

But when we look at the differences between marks—both average and largest differences—geometry is no better than history Here in a supposedly exact science we find nearly as wide differences as in history A part of the reason for this is that mathematics examiners throw away the advantage they have in more precise marking by using a method or scale of marking quite different from that used in any other subject For the *judgement* used in adjusting marks in other subjects the mathematicians substitute *simple arithmetic* For example, they ignore such things as the difficulty of the question, the expected ability level of the group etc , and award marks quite mechanically This results in a different scale for mathematics marks * (This is strikingly evident in Figure 4 2, and even more so in Figure 4 4) Comparing marks in biology and mathematics is like comparing degrees Fahrenheit with degrees Centigrade—

†† The highest mark observed in a sample tends to be an under-estimate of the corresponding highest mark in the population a comment which does not apply to the Mean or the S D of marks observed in the sample Gayen (1962) p 31

* We seem to be arguing for less judgement in history and more judgement in mathematics Let not the reader be confused by the apparent contradiction We are involved here with two different kinds of judgement One is the judgement on the basis of which candidates are placed in rank order (which is the bare essential of marking ') This judgement factor should be reduced and made as objective as possible. The second type of judgement however is the decision as to how many marks to award

and assuming that, both being "degrees", they should represent equal amounts of heat (This subject is dealt with in greater detail later, in the chapter on 'Scaling')

It is interesting to compare our one "highly experienced examiner" (Examiner X) with the average of the 6 examiners who re-marked his answer books In history, Hindi, and biology the mean marks of X and the mean of the six were almost identical In geometry, the six had average marks 2 01 marks below X's mean But in each subject there was more than half a standard deviation difference between X and the six, with the difference in history being one SD The differences, then, are in the number of high and low Divisions awarded, not in average marks

It is also interesting to ask whether there is any correlation between mean and standard deviation In other words, do examiners who "mark higher also tend to "distribute more widely", and vice versa? There *may* be some correspondence, but if so it is not high About 3 or 4 out of every 6 seem to agree fairly closely in rank on both measures, but 2 or 3 disagree so widely that they bring the correlation way down The rank order correlations range from + 54 to — 20, all of which are well within the range that *might* have occurred just by chance

Note that in Table 10 3 the Hindi marks are out of 34, while all others are out of 50 But in the last Hindi row, and in Figure 4 2 in order to make the 'pictures" comparable, Hindi figures have been multiplied by 50/34 to make them equivalent to marks out of 50 Thus we can see that the results for Hindi are really fairly close to both history and biology The Hindi averages are not quite so variable as history, but slightly more so than biology The Hindi ranges, however, are less variable than either biology, history, or even geometry Apparently the examiners of the Hindi Prose paper have a fairly clear idea of what their standard is As we saw from Tables 4 1 and 10 2 however, they often differ as to *where* within that standard a particular individual belongs

We have dealt here with averages, but the reader may also be interested in the detailed scatter charts given for each pair of examiners, in the following four chapters There, the full extent of inter examiner variability will be clearly evident

In summary, therefore we see that even *when six experienced examiners are given virtually identical batches of answer books, in history,*

a given rank It is here that *quite wrongly*, the mathematics teachers of India have abdicated their right of judgement while those of most other subjects have retained it The way in which judgement can be applied to the scaling of objectively-determined rankings is described by Gull ksen (1950) pages 265 266 (This has been adapted to Indian conditions by Harper (1963)) Of course, in certain circumstances there are purely objective non judgemental methods for scaling also

TABLE 10 3

FOR EACH OF THE SUBJECTS WE TOOK 300 ANSWER BOOKS OF ONE SPECIALLY EXPERIENCED EXAMINER AND DIVIDED THEM INTO SIX SETS WITH EQUAL (AS NEARLY AS POSSIBLE) MEANS AND STANDARD DEVIATIONS EACH SET OF 50 WAS RE MARKED BY A DIFFERENT EXAMINER THE FOLLOWING ARE THE RESULTS ARRANGED IN ORDER OF MEAN MARKS (SEE, ALSO FIGURE 4.2)

			Marks		Diff between marks	
			Mean	σ	Mean of differences	Largest differenc
HISTORY (50 marks)	Examiner X (standard)		18 10	3 814*		
		E	14 67*	4 899	3 92	10
		I	15 98	3 962	2 94	8
		F	19 46	5 100	2 84*	10
		J	19 82	5 133	3 04	11
		G	20 26	4 330	2 92	7
		H	22 06**	5 464**	4 52**	16**
	Range—h ghest to lowest		7 44	1 650		
HINDI (34 marks)	Exam ner X (standard)		11 84	3 742**		
		F	10 68*	2.672*	2.04	7
		I	11 84	3 540	1 44	4
		H	11 56	3.232	1 36*	7
		G	11 98	2 753	2 00	6
		E	12.34	3 439	1 42	4
		J	13 24**	3 314	2 32**	8**
	Range—highest to lo vest (Range out of 50 marks)		2.56 (3 76)	1 070 (1 574)	(3 41**)	(12**)
BIOLOGY (50 marks)	Exam ner X (standard)		20 42	4 540		
		F	19 34*	4 260	2 56	8
		E	20 04	7 411**	3 86**	9**
		I	20 06	5 270	2 36	7
		G	20 68	3 834*	1 70*	6
		J	20 74	5 114	2.34	7
		H	22 36**	5 309	3 10	9**
	Range—h ghest to lowest		3 02	3 577		
GEOMETRY (50 marks)	Exam ner X (standard)		31 19**	11 114		
		F	28 52*	10 828	3 18	14**
		H	28 62	12 665**	3 42**	13
		G	28 70	11 780	3 38	11
		I	28 94	11 860	2 64	14**
		E	29 24	12 046	3.26	11
		J	31 06	10 689*	2.52*	9
	Range—h ghest to lowest		? 67	1 976		

** H ghest or largest value in the set

* Lowest or smallest value in the set

Hindi, biology, or geometry, they differ widely, both in general and on individual cases

14 Does the number of pages written influence marks?

One old superstition can probably be laid to rest on the basis of these data

An American college magazine once published a picture of a student dressed as a professor, demonstrating the method of marking answer books The "professor was pictured standing at the top of a long flight of stairs He had tossed the entire pile of answer books down the stairs Obviously those with the largest number of pages were the heaviest, and reached the bottom These received the highest marks But the answer books with very few written pages were too light to travel far They stayed at the top of the stairs, and received the lowest marks Indian and American students share the conviction that many examiners just count the number of pages, and award marks accordingly

Is this true ?

Table 10 4 examines this question in detail The number of pages written in each answer book was counted and this number was correlated with the marks awarded, in both the first and second marking of each answer book Table 10 4 reports these product moment coefficients

First, we need some sort of a hypothesis Should we expect 'by chance" *no* relationship between the number of pages written and the number of marks awarded? This hardly seems reasonable, at least not for the extreme cases A student who knows very little will be unable to write more than a few pages, a student who knows a great deal will tend to write a fairly large number of pages to include all he knows But in between, of course, are the students who know a *little but 'pad it out* to fill many pages, and the students who know a fair amount, but write with brevity Thus, it seems reasonable to expect *some* relationship between number of pages written and number of marks awarded—but not a high relationship

This is just what the data show

We can say that an examiner is influenced by the number of pages if the correlation of his marks with pages is 'significantly higher than we should expect' To do this we need some base hypothesis as to the correlation expected In the absence of any better hypothesis it seems reasonable to take the median value of all correlation coefficients for a particular subject, as the "expected' or 'reasonable value for that subject Then we can examine whether any particular examiner s correlation coefficients differ significantly from that value

In Table 10 4 we find (*) to indicate 'significant" differences 15%

level of confidence), and (**) to indicate "highly significant" in the positive (i e. above expectation) direction. In other words, perhaps 10% of the time we have a suspiciously high correlation between pages and marks.

But let us look further.

(1) It seems reasonable to assume that if an examiner is influenced, he is likely to be influenced on all batches of answer books that he examines. (2) We have no evidence *against* the proposition that in a particular batch of answer books, there *may* be a higher "reasonable" relationship (due to the special quality of those candidates) than exists for the average.

Proposition (2) may be re-stated to the effect that correlation of pages with marks may be a function of the batch of answer books (rather than of the examiner). The bottom section of each of the four columns seems to support this hypothesis. Look at the pairs where Examiner X is paired with Examiner E, F, etc., through J. (Remember that 300 answer books of Examiner X were divided into six batches of 50 each, with approximately the same means and standard deviations. Also note, of course, that "Examiner X" is a different examiner for each of the four subjects. This is true of the other examiners also) Thus six samples of each Examiner X's answer books were re-marked by six *different* examiners. But a look at the pairs of marks will show that there is *definitely* a correlation between them. In the History column, for example, .52 is paired with .36, then .18 is paired with .12. *Thus we may conclude that correlation between number of pages and marks is frequently the function of the answer books themselves, rather than of the examiners.* (There is another possible explanation It just might be that both examiners tended to be highly influenced by one set of books, and not by another. This seems unlikely, when 7 examiners are involved in each subject, a total of 28 examiners for this part of the analysis).

So now we go back again to our "highly significant" figures. It seems evident that *some* of them may be due to the answer books rather than to the examiners. How can we tell? (1) We can see if any examiner is *consistently* too high (i e. significantly high for all his answer books). (2) We can compare pairs of examiners on the same answer books.

By the first test, we find only two examiners all of whose correlations are significant at the 5% level or better: Examiners C and J of history. (Only two other examiners come anywhere near "consistency" and both of them—Examiners C and E of mathematics—are consistently *independent* of the number of pages)

For the second test, the significance of the difference was calculated (using z) for all suspicious pairs of coefficients. Again, History Examiner C differed significantly from Examiner D on both batches of tests that were marked by both However, history Examiner J did not differ significantly

from Examiner X on the one batch of tests that both examined

It is interesting to note that the median *r* was lower for history than for any other subject Yet it is only in history that we have found any evidence of undue influence of number of pages on marks

So far we have been talking about *statistical* significance But in this case, psychological and educational significance seem to be accurately reflected by statistical significance (This is partly because N's are small, so *r*'s and differences need to be large to be statistically significant) Inspection of the figures suggests that few if any of the correlations that are not statistically significant could be expected to have any educational or psychological significance either, even if N s were large enough to make them statistically significant

We thus conclude that *only two out of 44 examiners were significantly influenced merely by the number of pages written by the candidate* This is less than 5% Many statisticans would, in fact, go a step further. They would say that if you find evidence of relationship in less than 5%, then this itself may be due merely to "chance" and not to any actual *relationship This certainly* would have been true if the significance had been only at the 5% level

The superstition that one gets a higher mark if one writes a lot, regardless of contents, may therefore now be laid at rest And it would be a boon to examiners if students stopped believing this old superstition, and did not fill up their answer books with so much rubbish!

11

Detailed Statistics—History

THE HISTORY paper consisted of 10 questions, out of which each candidate was to answer any five In one of the questions, the candidate had a choice of 5 out of 8 parts Each question carried 10 marks for a maximum of 50 marks for the paper

1 Distribution of marks

Table 11 1 gives the frequency distribution for each batch of 50 answer books in History These are the detailed frequencies on which the totals given in Table 4 5 are based

Table 11 1 A shows the frequency distributions for the first marking of each batch of answer books The top row names the original examiner of each batch It is his marks which are distributed in this part of the Table

The hundred answer books of Examiner A were split into two equivalent batches of 50 answer books each One of these batches was to be re marked by Examiner A himself, the other by Examiner B, as shown in the second row of the Table The answer books of Examiners B C and D were treated in the same way The reader can see, from this Table that the batches were as closely matched as was reasonably possible

The 300 answer books of Examiner X were split into 6 equivalent batches The matching here also is as close as is feasible

Table 11 1-B shows the marks awarded each batch of 50 answer books, the second time they were marked Compare for example the third pair of columns The two batches of Examiner C s work, which were

TABLE 11.2-B

FREQUENCY DISTRIBUTION OF *Second Examiner* TO EACH BATCH OF 50 ANSWER BOOKS IN HISTORY

Batch for re-marking by			E	F	G	H	I	J	X						Re-marking
Examiner	A	B	E	F	G	H	I	J	E	F	G	H	I	J	
50															
49															
48															
47															
46															
45															
44															
43															
42															
41															
40															
39															
38															
37													1		1
36													—		—
35															1
34												1	—		2
33												—	—		1
32												—	—		2
31	1									2		1	—		8
30	—				1			1				1	—		6
29	—				—			2			—	—	1		19
28	1	1			—	1		1			2	5	1		21
27	2	3			—	1	1	2			1	3	6		29
26	1	2			—	2	2	3			—	5	—		30
25	3	—			1	1	5	5			4	1	2		41
24	1	3			2	3	2	4		1	3	2	4		32
23	3	4			2	2	1				3	1			41
22	2	2			4	2	1	3		—	3	5	2	2	59
21	5	6			3	6	4	1		4	2	6	4	5	60
20	6	2		3	4	—	4	5		4	2	6	7	6	59
19	—	4		2	2	5	2	1		5	3	7	9	2	81
18	2	5		5	4	10	5	4		3	6	2	7	5	125
17	14	5		—	6	6	4	10		3	7	2	9	5	20
16	—	—	2	—	—	1	—	1	2	—	6	4	—	—	69
15	—	2	7	4	2	6	3	7	5	1	1	1	1	4	68
14	2	3	6	2	3	1	2	5	5	2	1	2	1	2	37
13	1	2	1	3	1	1	3	2	4	1	2	—	1	1	41
12	2	1	5	1	4	—	4	6	1	3	—	—	—	1	39
11	1	2	1	4	4	—	3	4	3	1	1	—	2	2	25
10	—	1	5	3	3	—	1	2	1	2	—	—	1	1	36
9		1	3	2	—	—	—	—	3	—	—	—	—	2	11
8		—	2	—	2	—	—	1	1	1	1	1	1	2	15
7		1	—	1	—	—	1	—	2	—	—	2	—	1	5
6		—	1	—	—	—	—	—	1	1	—	2	—	1	4
5			—	1	—	2	—	—	—	—	—	1	—	2	6
4			1	—	—	—	1	—	—	—	—	—	—	—	1
3			—	—	—	—	—	—	—	—	—	—	—	—	3
2			—	—	—	—	—	—	—	—	—	—	—	—	1
1			—	—	—	—	—	—	—	—	—	—	—	—	1
0			1	—	—	—	—	—	1	—	—	—	—	—	1
Number	50	50	50	50	50	50	50	50	50	50	50	50	50	50	1000

— is used in a column to mark, when there were marks awarded above or below this This is to emphasize certain instances

closely matched in the first part of the Table, are now awarded quite different distributions of marks by Examiners C and D

The lines across the Tables show the divisions between the Classes or Divisions Notice the very striking J-effect in the first marking (11 1-A) A large number of candidates were awarded 17 (passing), and *no* candidates were awarded 16 In the re marking (11 1 B), however, quite a few 16 s were awarded—though the J effect by no means disappears

There is also a slight J effect between First and Second Divisions in the first marking—but, again, not in the second And there is an interesting reverse J between II and III in both markings Apparently history examiners are a little reluctant to give a Second Division—so they tend to pile up candidates with 22 marks, rather than award a 23

Note that only 37 of the 51 points (50+0) in the 'scale' are actually used The marks from 37—50 are wasted—they might just as well not exist

2 Table 11 2

Table 11.2 summarizes most of the statistical tables produced by the IBM "1620" computer The summary means and medians of these statistics have already been reported in Table 10 1

The top row tells who originally examined the batch of 50 history answer books which provided the statistics in the column The second row tells who examined the answer books the second time The third row merely notes (for easier reference) whether the answer books were re examined by the Same examiner, or a Different one from the original

In the remainder of the Table, "O" refers to the Original marking, and "R" to the Re-examining

3 Mean and standard deviation

The wide variety of means and SD's should perhaps not be taken too seriously Though the samples of the work of each examiner are not believed to be biased in any significant way, they are also not strictly "random" The exigencies of the situation required that we take the first hundred answer books that we could find to represent each examiner (see Chapter VI) There is no reason to believe that they were essentially different from the rest of the work of that examiner in any manner significant to this study However, we cannot assume that they *exactly* resemble the mean and standard deviation of his whole batch This fact does not affect the differences between marking and re-marking, which are the prime consideration of this study

The signs attached to the row of Differences between the means are

arbitrary. When the second mean is higher, we have attached a "+"—when the mean for re-examining is lower, we have attached a "—". This seems logical (or at least psychological!).

The same comments apply to the SD's (standard deviations). A"+" means the second examiner widened out the distribution, a "—" means that he attenuated, narrowed or compressed it.

Note, however, that in a sense the "direction" is really irrelevant. "First" and "second" examiners could easily be interchanged. There are some consistent features of the "second" examiners—on the average they have marked a little lower, and with less J-effect. However, the far more important factor is the actual differences, regardless of "direction".

Differences between means range up to 5.02 (which is 10% of marks). Differences in SD range up to 3.061 (or 6% of marks).

For the difference between means, *indicates that the difference is significant at the 5% level of confidence; **that it is significant at the 1% level of confidence; and ***that the difference is significant at the 0.1% level of confidence.

The interpretation for the asterisks attached to the SD's is the same, except that, unfortunately, we had no tables for significance at the 0.1% level of confidence here.

We may summarize these data as follows, giving the number of differences which are significant at each level of confidence:

	.001	.01	.05	NS
M	13	4	1	2
SD	××	6	3	11

	Re-examiner Same	Diff.
Neither M nor SD differences are significant	0	0
Mean difference only is significant	7	4
SD difference only is significant	1	1
Both M and differences are significant	2	5

Remember that if the examiners are *reliable* (in the broad sense of the term), then the differences should be *not* significant. *No batch of papers was marked by a pair of examiners whose differences were small enough to be attributed only to random chance.* All errors were *significant* errors. indicating that factors other than sheer chance were operating

4. Distribution of Divisions

Those who do not understand means and standard deviations will get a clear picture of what is happening when they look at the next few

rows These show how many candidates were awarded each Division by each examiner This is a good way of knowing the subjective standard that the examiner has in mind when he marks the answer books

In each major column you will find two columns of figures in this section These show the number (out of 50) awarded each Division the first and second time i e by the Original and Re examiner

For example Examiner A awarded only one First Division the first time he marked these 50 answer books but awarded three Firsts when he re-examined the same answer books Similarly he failed 9 the first time but enlarged this to 15 the second time It is the larger number of Firsts *and* Fails which is responsible for the larger SD on re marking for Examiner A

In almost all cases *there was a wide difference in the number of candidates awarded each Division by the same examiner the first and second time he marked his answer books*

The difference was even greater when it was a different examiner who re-examined the answer books This can be seen in the second fourth sixth etc columns Examiner H for example awarded four times as many Seconds and Firsts as did Examiner X Examiner D failed nearly 7 times as many as Examiner C

5 Pass percentage

A popular (though very crude and unreliable) measure of quality is the Pass Percentage A glance along the next three rows shows how little trust can really be placed in this figure The first of these three rows shows the percentage passed in the Original marking the second row the percentage passed in the Re-marking and the third row the difference between these two percentages

Notice that even the same examiner re examining his own answer books differed from himself by as much as 36% in Pass Percentage (Examiner D) No examiner s pass percentage remained exactly the same though Examiners G and J came close

Examiners X and F surprisingly enough agreed perfectly in Pass Percentage—though other parts of this Table show that they agreed on nothing else But other differences went as high as 46% (Examiners C and D)—though when it was C re marking D s answer books instead of vice versa the difference was only 26% This latter fact again shows the lack of self consistency of examiners

6 Examiner reliability

There is a wide variation among the correlations between each pair of marks

The self-correlations of examiners re-examining their own answer books ranged from 72 (Examiner I) to 93 (Examiner F) The median was 85, indicating a fair degree of agreement This compares favourably with such studies done in other countries

But when examiners marked each other's answer books, the product-moment coefficient of correlation dropped as low as 49 (D—C) The highest correlation between two history examiners was 85, which is the same as the median self-correlation

It will be remembered that the previous study of 90 examiners marking the same 10 history answer books, estimated the average rank order inter-correlation as 83 Why does the present study show somewhat lower inter-correlations (median= 74)? We do not know One possible reason may be that, while in the first study we selected only answer books in which all candidates had answered the same five questions, in this study there was no such restriction It is well known that allowing a choice of questions reduces the reliability and validity of an examination It is possible that it may reduce reliability of marking, also

But even the highest self-correlation, .93 for Examiner F, does not look so good when you examine the differences in M and SD (1 08 and 413)—or the differences in the Divisions and Pass Percentage (12%)—for this examiner This clearly illustrates the fact that the correlation coefficient alone is misleading, as it takes no account of differences in M and SD

7 Content reliability

The calculations for content reliability (see Chapter VII) were done on the basis of the original marks awarded each batch of answer books *

* An interesting exception to this must be noted in the case of examiner D The calculated content reliability for this examiner s 100 answer books is —1 182 Obviously this is absurd being both negative and more than unity Furthermore the examiner reliability of 494 indicates that there must be at least a minimum of content reliability On investigation this odd result was found to be due to the abnormally small SD (about 2 marks out of 50) for this examiner We therefore based our estimate of content reliability on two other calculations

D re-marking his own answer books	.259
C re-marking D s answer books	775

(Note that C's result here is slightly higher than the reliability of his own first marking 755—though the difference may not be significant) These two reliabilities were translated into Fisher s z averaged and the mean z translated back into r— 570 The skeptical may think this negative reliability invalidates the formula used Not so An objective examination of 36 questions (equivalent to the range of marks actually used by history examiners) a mean of 18 74 and a SD of 1 96, would have a Kuder Richardson Formula 21 reliability of —1 38 (If we take number of questions as 50 the reliability becomes —2 09) These cases of negative reliability above unity would not lead us to throw out KR 21—we would only throw out the examination!

(Note that the batches for A B C D are 100 each for E, F, G H, I J are 50 each, and for Examiner X the number of answer books is 300)

The fact that there is a wide variety of content reliabilities for batches of answer books from the *same* examination needs some explanation In the first place, some of these differences may be due to *different choices* of questions However probably more important is each examiner s style or method of marking Thus two examiners may expect two different kinds of answers (not completely different, of course but different in significant aspects) from the same question In a sense therefore we do not really have a single examination—*we have as many examination papers as we have examiners* So students sitting for the same examination are really sitting for different examinations—and because the student knows nothing about his examiner, *he does not know which examination he is sitting for* This is in line with Taylor s remark (which repeats in India similar studies done in the West), that the candidate s mark depends more on the luck of who his examiner is than on his actual merit.

Content reliabilities range from 21 (for H) to 76 (for C) But even this highest value would be considered grossly inadequate in an objective examination

8 Marks reliability

As has been explained earlier (Chapter VII) the only fair comparison between traditional and objective examinations is on the basis of what we have termed "marks reliability This is the product of examiner and content reliability We have calculated this only in the cases where a different examiner has re marked the answer books ' *Marks reliability* ' attempts to estimate the correlation between the marks of a batch of candidates who have answered examination I marked by examiner P, and examination II marked by examiner Q

Marks reliabilities range from a low of 28 to a high of 43 Even this highest reliability is not high enough to estimate the *class or group average* adequately—let alone estimate the true merit of any individual

Of course, the reliability of the history mark based on *two* papers *is somewhat higher than* the reliability of a single paper However, the difference is not enough to justify the heavy reliance placed on history marks

In a sentence *History marks are unreliable*

9 Standard error of measurement (SEM)

The SEM ranges from 4 48 to 3 01 This means that even for the most acurate batch, there are only two chances in three that a given

candidate's mark will be within 3 out of 50 of his "true" mark. For about 5 candidates in every 100, the error will be 12% of marks or more, for this best batch. But as has been pointed out, our concern should be for the largest (not the smallest) errors that can occur. A SEM of 4.48 means that for one candidate in each hundred (or 1,000 in every lakh), the error will be 23% of marks or more. And whether the mark awarded will be 23% *above* or 23% *below* the "true" mark is a matter of sheer chance.

But note that these estimates of errors take into account only the correlation—not the differences in M and SD. Thus they would be true only if the marks of all examiners were statistically scaled to a common standard. More exactly, each SEM shows the range of error which would exist if the re-marking by the second examiner were scaled to the mean and standard deviation of the original marking of the original examiner. The actual range of error, under current unscaled examination marking methods, would probably be much greater (see discussion in Chapter VII).

10 Differences

The last section of Table 11.2 presents the frequency with which each difference between examiners of a single answer book appeared. Note that this does *not* take into account the content reliability (as the SEM does), but only the actual difference between two markers of the same answer book. If each candidate had answered two parallel examinations, marked by different examiners, these differences would be much larger.

The largest difference for the same examiner re-examining 50 of his own answer books was 10 marks. There were three such differences. The details of one of them (Examiner C) are as follows:

Question	First Marking	Second Marking
1	4	6
2	2	4
3	5	8
4	6	8
9	2	2½
	19	28½=29

N Examiners in Batches of 50 Answer Books Each

X	H	X	I	X	J	X
G	H	H	I	I	J	J
Diff	S	Diff	S	Diff	S	Diff
18 10	15 42	18 10	16 58	18 12	17 46	18 10
20.26	17 56	22 06	15.22	15 98	19 32	19 82
+ 2.16***	+ 2.14***	+ 3 96***	− 1 36***	− 2 14***	+ 1 86***	+ 1 72**
376	423	578	341	397	403	515
5 745	5 059	6 851	3 988	5.390	4 615	3.340
3 915	5 154	3 748	2.967	3 734	4 424	3 885
4 330	5.525	5 464	3 419	3 962	5 795	5 133
+ 415	+ 371	+ 1 716**	+ 452	+ .228	+ 1 371**	+ 1.248*
1		3			1	1
5 12	5 11	6 21	2	6	7 18	5 14
29 29	20 20	32 18	36 22	32 31	26 13	33 22
15 9	25 19	12 8	12 28	12 19	17 18	12 13
70%	50%	76%	76%	76%	66%	76%
82%	62%	84%	44%	62%	64%	74%
12%	12%	8%	32%	14%	2%	2%
797	845	665	723	735	878	706
478	.213	478	.286	478	487	478
381		318		351		337
3 08		3 10		3 01		3 16
		1				
		—				
		—				
		—				
		1				
		2				1
						2
		3				
	3	2		1	1	2
	2	3		2	1	2
1	1	3	2	2	1	2
3	3	8	4	5	8	2
8	10	5	4	9	7	5
5	9	3	11	10	9	5
12	7	9	9	4	7	10
8	7	7	13	14	9	16
9	8	3	7	3	7	3
4						
2 92	2 94	4 52	2 24	2.94	2.78	3 04

F)

The largest difference for a pair of examiners was between Examiners X and H, and was 16 marks The details are as follows

Question	First Marking	Second Marking
1	4	8
2	3	6
3	4	8
4	3	6
10	4	6
	18	34

Since this largest difference occurred once in 500 answer books the probability is that there would be "errors" of 16 or more for at least 200 out of every lakh of candidates

Note that by chance the two examples chosen are both examples of the second examiner raising the marks But nearly as often, the second examiner lowered the marks

It is also instructive to look at a case, chosen nearly at random, of "agreement" between two examiners

Question	First Marking	Second Marking
1	5	5
2	2½	3
3	3	4
4	3	4
9	4	1½
	17½	17½

We will let the reader draw his own conclusions from these marks

The average differences are reported at the bottom of each column in Table 11 2 The smallest is 1 68 (Examiner F), and the largest 5 14 (Examiners C and D) Thus they range from 3% to 10% But putting much faith in such averages reminds one of the man who wanted to go and live in the desert When his friends remonstrated, he admitted that the days would be hotter than 120° and the nights colder than 35°—but he was sure that the *average* of 77° would be just wonderful ! Realistically, it is the large differences which are possible that should concern us, not the averages Even the Divisions of forty per cent of the candidates

were different in the two markings There is little comfort in the "average", when the examination can be so very unjust to so many

11 Scatter charts

Table 11 3 compares the work of Examiners A and B on their own and on each other's answer books It also compares the re-marking of examiners C and D with each other

These and Table 11 4 are the individual scatter charts which are summarized in Table 4 3

For example, look at the first chart, showing what happened when Examiner A re examined his own 50 answer books Out of 29 that he awarded a Third Division the first time (line 3 and column T, which means Total), the second time he reduced 7 of them to Fail, raised 6 to a Second Division, and awarded one a First Conversely, of the three who were given a First Division in the second marking, only one had been a First originally, and the other two were a Second and a Third Yet this is an examiner with the relatively high self-correlation of 89

The little dashes (—) indicate zero cases at the points which would show perfect agreement on Division They are to help the eye to find the diagonal of perfect agreement, in each chart

At the bottom of the chart we also find the words, "Agree=64%' This refers to the fact that 64% of the candidates received the same Class or Division in both markings by Examiner A

The remaining charts may be read in the same way Note that the top row of four charts in Table 11 3 show the self agreements of the four examiners, the bottom row shows their agreements with each other.

In the bottom row, it is interesting to note that Examiners C and D seem to agree with each other much less than examiners A and B agree This may be partly due to different styles or concepts of marking Especially note the *very narrow* distribution of Examiner D's marks, both times with his own answer books, and also when re examining Examiner C's answer books Where C distributes his groups from I to Fail, D uses only two categories

This is a rather striking illustration of the fact that the distribution of marks (and, hence, number in each Division) is often more a characteristic of the examiner than of the batch of answer books examined For the *same* batches, C distributes widely, D narrowly Thus the student's Division depends more on luck than on merit It might also be added that the lower self-correlations of C and D suggest that they may be somewhat more careless (less self-consistent) than Examiners A or B *

* The reader may be tempted to think that D s lower self-correlation is due to ' restriction of range ', since he uses only two categories But the ' restriction of range effect '

TABLE 11 3

PAIRS OF EXAMINERS MARKING THEIR OWN AND EACH OTHER S CANDIDATES IN *History*

OWN

Second Examiner—*A*

First Examiner—*A*	F	3	2	1	D	T
D				1	—	1
1			2	8	1	11
2	7	15	6	1		29
3	8	1				9
T						
T	15	18	14	3		50

Agree = 64%
r = 89

Second Examiner—*B*

First Examiner—*B*	F	3	2	1	D	T
D					—	
1			1	5		6
2	8	12				20
3	22	2				24
F						
T	30	15	5			50

Agree = 78%
r = 86

Second Examiner—*C*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1			1	—		1
2		1	7	3		11
3		16	17			33
F	2	3				5
T						
T	2	20	25	3		50

Agree = 50%
r = 78

Second Examiner—*D*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1			1	—		2
2	18	29				47
3	1					1
F						
T	19	30	1			50

Agree = 62%
r = 74

OTHERS

Second Examiner—*B*

First Examiner—*A*	F	3	2	1	D	T
D					—	
1			2	—		2
2	17	8	1	1		10
3	8	12				29
F		1				9
T						
T	25	21	3	1		50

Agree = 42%
r = 80

Second Examiner—*A*

First Examiner—*B*	F	3	2	1	D	T
D					—	
1			1	4		5
2	3	11	8			22
3	19	4				23
F						
T	22	16	12			50

Agree = 68%
r = 85

Second Examiner—*D*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1		2		—		2
2	6	5	—			11
3	17	16				33
F	4					4
T						
T	27	23				50

Agree = 40%
r = 50

Second Examiner—*C*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1				—		
2			3			3
3	13	24	10			47
F	—					
T						
T	13	24	13			50

Agree = 54%
r = 49

AA—AB = z = 32
BB—AB = z = 18

AA—BA = z = 17
BB—BA = z = 01

CC—CD = z = 50*
DD—CD = z = 41*

CC—DC = z = 51*
DD—DC = z = 42*

* A difference of 40 is significant at 5% level of confidence

** A difference of 53 is significant at 1% level of confidence

Do examiners agree with themselves any better than they agree with each other? All r's were translated into Fisher's z, and the differences calculated. The last two lines of Table 11.3 give the result. Examiners A and B seem to agree with each other just as well as they agree with themselves. But examiners C and D agree significantly* better with themselves than with each other.

The left half of Table 11.4 shows the self-correlation of Examiners E, F, G, H, I, J. Note that even Examiner F, who had the highest self-agreement (+.93), spread both his II and III candidates over three Divisions each, the second time he marked them.

Percentage agreement and r are, of course, different measures of agreement. One reason why they are not closely related is that a perfect correlation of marks can exist, even if the Division are entirely different. Another reason is that the Division categories are of unequal size. 'Fail' covers a range of 17 marks (including 0), 'III' only 6.

The right half of Table 11.4 shows what happened to the six equivalent batches into which Examiner X's 300 answer books were divided. If you look at the Totals in the extreme right column of each chart, you will see that the various batches did differ slightly. (If there was only one First Divisioner, he could not be assigned to all six re-examiners.) However, look at the Total row at the bottom of each scatter chart, comparing all six of them with each other. Examiner I awarded only III and Fail. (Note that his own answer books, in the left hand column of scatter charts, show an equally narrow range.) In contrast to this, Examiners F, H, and J, marking virtually identical batches of answer books, spread them from Fail to First. Other examiners fall between.

Examiner I, surprisingly, agrees with Examiner X slightly better than he agrees with himself. The difference is not statistically significant—the point is that he does equally well whether marking his own or some one else's answer books.

Examiner J shows a higher percentage of identical Divisions (64% vs. 58%) but a signficantly lower product-moment correlation (.71 vs. .88), with X than with himself.

is due to truncation of a distribution by elimination of cases at one or both extremes. In this case there is no truncation, only a forcing of an extremely leptokurtic (and also badly skewed on the first marking) distribution on the data.

* Two-tailed test of significance. At first it seemed logical to hypothesize that examiners agreed better with themselves than with each other. This would have allowed us to use the more powerful one-tailed test of significance. However, examination of the data showed so many instances where self-agreement was *lower* than other-agreement (see subsequent chapters, especially mathematics) that we had to discard our original hypothesis, and adopt the two-tailed test. Any reader who prefers to apply the one-tailed test may do so by noting that .33 is significant at the 5% level and .45 at the 1% level.

Out of fourteen comparisons between correlation with self and with another examiner, the differences for one-half (i.e. 7) were significant, two of them being highly significant. In only one instance was the correlation with Other higher than with Self, and that was not significant. This is one more bit of evidence that each history examiner has his own special style of examining, and that there is not a high degree of agreement among examiners as to what is and is not important.

The interested reader may make other detailed comparisons between scatter charts for himself.

12. Self-consistency of examiners' distributions

We have already discussed the product-moment correlation as a measure of self-consistency. Another measure is the extent to which examiners tend to agree with themselves in M, SD, and percentages of marks in each Division.

We get a more critical picture of this when we examine each pair of examiners. Take, for example, Examiners A and B. Remember the experimental design: The hundred answer books originally marked by Examiner A were split into two equivalent batches of 50 each. Similarly, B's 100 answer books were split into two equivalent batches. Fifty of Examiner A's and 50 of Examiner B's answer books were then mixed up in a random order, and sent to Examiner A for re-marking. Similarly, B received a mixed batch of both his own and A's answer books. The examiners were told only that "some of the books" would be answer books originally marked by themselves, some originally marked by others. In response to a questionnaire, examiners assured us that they were generally not aware of which books were which.

Here is the comparison of A and B (from Table 11.2):

	A	B		A	B
	19 1	15 5		5 30	5 24
Mean			SD		
	16 9	15 8		6.36	5 38

In the first row, the 50 answer books to which A originally awarded mean marks of 19.1 were awarded mean marks of 15.5 by Examiner B. The second row must be read backwards. The batch to which B originally awarded a mean of 15 8, were awarded mean marks of 16 9 by A. Thus for both batches, B's mean is consistently lower than A's. Similarly, B's SD is consistently smaller than A's.

Here is the same comparison for C and D, who were also a matched pair in the experimental design

	C	D		C	D
	19 9	14 9		4 00	3 35
Mean			SD		
	18 9	18 7		5 02	1 96

The difference between C s and D s ' styles of marking is even more striking if we look in Table 11 2 and in the scatter charts at the percentages each assigns to each Division

For more detail let us compare the six examiners who were matched against nearly identical batches of Examiner X s marking

	Means when re-marking		SDs when re marking	
	Self	X	Self	X
E	12 7	14 6	4 49	4 90
F	16 8	19 5	4 97	5 10
G	18 8	20 3	4 64	4 33
H	17 6	22 1	5 52	5 46
I	15 2	16 0	3 42	3 96
J	19 3	19 8	5 80	5 13

Note that, with the exception of a reversal of H and J, the rank orders are identical in the two columns of Means (*rho*= 78) The rank order correlations of the SDs is 89

Also look in Table 11 1—B at the distributions of marks awarded matched groups of answer books Note, especially, D re marking C, and I re-examining X s answer books

All of these facts suggest that examiners *differ fairly consistently* from each other in ' standard of marking Each examiner applies his own concept of what the average marks should be how many should fail how many should be in each Division, to each batch of answer books that he marks Apparently the present instructions and "model papers , etc , issued do not serve to control these differences in "standard adequately *It would be wise for the authorities to specify more exactly the distributions which they expect for 'average or typical ' groups at least* (This is not a radical suggestion It is in fact common practice in many countries)

13 Item analysis

One of the most revealing procedures that any examiner can carry out is an item analysis At the simplest level, it takes very little work

Merely select a random group of answer books (say 100), arrange in order of merit, divide them into several equal groups, and then tabulate how well each of these groups did on each question

We divided our thousand History answer books into five groups of 200 answer books each, on the basis of the first marking Table 11 5 presents the results

We had agreed to keep secret the identity of the Board supplying these data Unfortunately, this precludes our publishing the history question paper However, its organization can be revealed, as it is fairly typical

There were 10 questions, each question carrying 10 marks The candidate was to answer any 5 out of the 10 Maximum marks were thus 50 Question 9 had 8 parts, of which the candidate (if he chose Question 9) was to answer any 5 parts, for 2 marks each

Table 11 5 gives the number of candidates of each ability level choosing each question, and their mean marks on that question For example, the average marks of the middle group were 17 37 (see right-hand column) One hundred and ninety-seven of these 200 candidates chose to answer Question 1 The average marks of these candidates on Question 1 were 4 34 On the other hand, candidates of equal ability (as judged by total marks) who chose Question 7 were awarded an average of only 1 96 marks Thus the average ability candidate lucky enough to choose Question 1 rather than Question 7 is awarded (on the average) 2 4 extra marks—and even the candidate choosing Question 9 gets 2 3 extra marks for his choice

Gayen (1961, pp 51-53) has emphasized the fact that a candidate's total marks are determined partly by his choice of questions This can be illustrated for the top-scoring group (top row of Table 11 5), by dividing the 10 questions into two sets of ' choices ' of five each

Question	Marks	Question	Marks
9	5 8	2	4 0
10	5 1	4	4 0
1	4 9	6	4 0
3	4 5	8	3 6
7	4 4	5	3 2
	24 7		18 8

Thus candidates of *equal* ability can get *different* marks just by happening to choose one or another set of questions We have argued elsewhere that the examiner who allows a choice of question, is stating

TABLE 11 5

ITEM ANALYSIS OF *History* HIGHER SECONDARY EXAMINATION (FOR DESCRIPTION OF EXAMINATION, SEE TEXT)

Question Number	1	2	3	4	5	6	7	8	9	10	Mean Marks
Mean marks of group on question	4 87	4 00	4 55	3 96	3 17	3 96	4 36	3 60	5 82	5 10	23 464
(Number in group choosing question)	(190)	(88)	(184)	(143)	(32)	(45)	(53)	(10)	(178)	(84)	(200)
Mean marks of group on question	4 39	3 30	3 94	3 25	2 95	3 44	3 03	2 00	5 04	3 21	19 665
(Number in group choosing question)	(197)	(94)	(193)	(124)	(31)	(72)	(46)	(9)	(165)	(80)	(200)
Mean marks of group on question	4 34	2 97	3 66	2 93	2 49	3 22	1 96	2 22	4 26	2 44	17 372
(Number in group choosing question)	(197)	(90)	(195)	(122)	(46)	(53)	(67)	(9)	(143)	(96)	(200)
Mean marks of group on question	4 48	2 91	3 62	2 82	2 18	2 65	2 29	2 36	4 08	2 10	16 118
(Number in group choosing question)	(181)	(87)	(189)	(144)	(50)	(68)	(75)	(11)	(119)	(87)	(200)
Mean marks of group on question	2 73	1 64	2 88	2 00	1 19	1 43	0 97	4 33	2 24	1 30	11 062
(Number in group choosing question)	(188)	(98)	(189)	(102)	(59)	(78)	(67)	(6)	(90)	(100)	(200)

that he believes the questions to be of equal *difficulty*. Since this is very obviously *not* true, and since the candidate has no way of knowing which questions the examiners will mark "stiffly" and which "easily", the traditional essay type examination ends up as a sort of guessing game. The student who is lucky enough to make the right guesses gets higher marks than he deserves—the poor fellow who guesses wrong suffers unjustly.

Note, in Table 11.5, that the questions differ quite widely in "popularity." Hardly any candidates selected Question 8—practically all of them selected Questions 1 and 3.

Most of the questions show a fair degree of relationship with total marks. But Question 8 is an excellent example of *negative* validity—the poorer the student (as judged by total marks), the better he does on Question 8.

Question 8

	Mean Marks	Average Marks per Question	Marks on Question 8
Top group	23.464	(4.7)	3.6
Second level	19.665	(3.9)	2.0
Third level	17.372	(3.5)	2.2
Fourth level	16.118	(3.2)	2.4
Poorest group	11.062	(2.2)	4.3

The students who averaged only 11 marks at the examination did much better on this question than those who averaged 19.7 marks (4⅓ marks vs 2). Thus the effect of Question 8 was to raise the marks of poor students, and lower the marks of the best, for those who chose it. The same effect is partly evident in Question 7, which shows a non linear relationship. This is just one more of the ways in which allowing a choice of questions lowers the validity of an examination.

A more sophisticated method of item analysis might have revealed these effects in more precise detail. This was not done as this was not a central concern of this study. However, as a side light, the results seem sufficiently revealing to justify the effort put into them.

12

Detailed Statistics—Hindi

THE HINDI paper consisted of 6 parts, all compulsory, with a wide selection of questions within each part This paper covered prose text grammar, unseen and supplementary books It was a three hour paper, and maximum marks were 34 In some of the following we have reported in terms of 34 marks, in other places in order to compare with the other three subjects we have multiplied marks by 50 34 to make them equivalent to the 'out of 50 marks' scale

1 Distribution of marks

Table 12 1 gives the frequency distribution for each batch of 50 answer books, in Hindi These are the detailed frequencies on which the totals given in Table 4 5 are based

Table 12 1 A shows the frequency distributions for the first marking of each batch of answer books The top row names the original examiner of each batch It is his marks which are distributed in this part of the Table

The hundred answer books of Examiner A were split into two equivalent batches of 50 answer books each One of these batches was re marked by Examiner A himself the other by Examiner B as shown in the second row of the Table The answer books of Examiners B, C, and D were treated in the same way The reader can see from this Table, that the batches were as closely matched as was reasonably possible Some adjustments had to be made to make the means and SDs equal.

The 300 answer books of Examiner X were split into 6 equivalent batches The matching here, also, was as close as is feasible.

Table 12 1-B shows the marks awarded each batch of 50 answer books, the second time they were marked Compare, for example, the third pair of columns The two batches of Examiner C's work, which were closely matched in the first part of the Table, are now awarded quite different distributions of marks by Examiners C and D A similar effect is evident in B's work when re evaluated by A and B

The lines across the Tables show the divisions between the Classes or Divisions * Notice the striking J-effect in the first marking (12 1-A) A large number of candidates (194) were awarded 11 (passing) but only a quarter as many (52) were awarded a failing 10 In re marking (12 1-B) this J-effect is greatly reduced, but by no means disappears Notice the very striking effect in Examiner D s answer books While the first time there were 16 11's and *no* 10 s in each batch, on the second marking more 10's than 11's were awarded ! Examiner H, on the other hand,

* The distribution for Hindi is interesting The maximum marks in this paper are 34 Thirty three per cent marks is of course, passing This is 12 marks Yet the J-effect shows that the examiners are obviously thinking of *eleven* (rather than twelve) marks as passing Thus they are apparently thinking of 32% marks as being the passing level on this particular paper

How can we explain this paradox ? It seems to be in the distribution of marks in the three papers, and the fact that it is the total marks on the three—not marks on any single paper—which determines passing The maximum marks on the three Hindi papers are 34 34, and 32 Note that 33% of these is 11 22 11 22, and 10 56 respectively Thus if 11 is considered passing on each paper, the total of 11+11+11=33 Hence the examiner who thinks, ' this candidate should just barely pass" will give him 11 marks (32%) rather than 12

This ' brinkmanship —which is typical of other subjects also—seems rather naive to say the least It seems to depend on the assumption that the ' should barely pass" candidate will receive the *same* marks on all papers Yet the probability of this is rather low, as various studies have shown that the correlation between papers is nowhere near perfect So why not do away with this—and abolish the J-curve in marking altogether ? The logic is simple The examiner should, we believe, reason as follows Only a few candidates will receive the same mark on all papers For most, the marks will *differ* So (in Hindi) instead of raising a 10 to an 11, why not give him 10 ? If he gets higher marks on the other papers, he will pass anyway, whether I give him 10 or 11 If he fails on the other papers, he will fail anyway, regardless of whether I give him 10 or 11 In the unlikely event that he gets just exactly 11 on each of the other two papers, of course be will fail if I give him only 10 But is there really any educational logic in passing him by giving him an extra unearned mark on Paper 1, when his total performance is so low anyway?" (Incidentally, India seems to have *about the lo west pass marks in the world* Some years ago the senior author had the privilege of teaching a class at the International Statistical Education Centre, in Calcutta. They represented some *dozen* Asian countries Once he asked them about the pass marks in each of their countries They ranged from 60% in the Philippine down to 33% for India and Pakistan—who shared the ' honour" of being at the bottom)

TABLE 12 1-B

FREQUE BY *Second Examiner* TO EACH BATCH OF 50 ANSWER BOOKS IN HINDI

Batch for marking Examine C	E	F	G	H	I	J	X						Re marking
	E	F	G	H	I	J	E	F	G	H	I	J	
M/34 M/													
34 5													
33 4													
32 4													
31 4													
30 4													
29 4													
28 4													
27 4													
26 3	1												1
25 3	—												—
24 3	—												—
23 3	1		1						1				3
22 3	—		1						—				1
21 3	—		—						—				1
20 2	2		1						—				5
19 2	2		1						—		1	3	9
18 2	—		—				2		1	2	—	5	14
17 2	6		4				3	1	—	1	2	3	30
16 2	2	3	8		1	1	5	2	1	4	1	2	38
15 2	6	1	12	1	5	2	4	1	2	2	7	—	55
14 21	4	5	5	4	4	—	6	3	7	3	9	7	87
13 13	5	3	5	2	8	7	5	6	7	5	4	12	105
12 15	5	10	3	6	9	5	8	5	11	11	6	4	121
11 15	9	7	4	10	9	13	6	7	5	6	8	4	134
10 18	1	5	2	—	3	9	2	8	7	—	1	4	77
9 17	2	5	1	6	4	1	1	6	4	8	2	3	87
8 14	1	4	2	10	3	5	2	6	3	6	1	1	84
7 14	1	2	2	4	2	4	2	2	—	—	2	1	46
6 4	2	1		4	—	2	2	2	1	—	4	—	47
5 4		3		2	1	1	2	1		—	1	—	29
4 —		—		—	1	2	—			—	—	1	13
3 4		—		1	1					2	—		9
2 1		1									—		2
1											1		2
0													
Number 0	50	50	50	50	50	50	50	50	50	50	50	50	1000

"—" is usmark, when there were marks awarded above or below this
This in certain instances

maintains a consistent J in both his original marking (Table 12 1 A) and in his re marking of both his own and X's answer books (Table 12 1 B)

There is also a slight J effect between the First and Second Divisions in the first marking—but again not in the second (The identical effect was noted in history also) And there is also (as in history) an interesting reverse J between II and III in both markings Apparently Hindi examiners like history examiners are a little reluctant to give a Second Division—so they tend to pile up candidates with 14 marks rather than award a 15

Note that only 26 of the 35 points (34+0) in the scale are actually used The marks from 27 to 34 are wasted—they might just as well not exist In a rational measurement scale full marks should be considered attainable

2 Table 12 2

Table 12 2 summarizes most of the statistical tables produced by the IBM '1620 computer The summary means and medians of these statistics have already been reported in Table 10 1

The top row tells who originally examined the batch of 50 Hindi answer books which provided the statistics in the column The second row tells who examined the answer books the second time The third row merely notes (for easier reference) whether the answer books were re marked by the Same examiner, or a Different one from the original

In the remainder of the Table O refers to the Original marking and R to the Re examining

3 Mean and standard deviation

The wide variety of means and SDs should perhaps not be taken too seriously Though the samples of the work of each examiner are not believed to be biased in any significant way they are also not strictly 'random The exigencies of the situation required that we take the first hundred answer books that we could find to represent each examiner (see Chapter VI) There is no reason to believe that they were essentially different from the rest of the work of that examiner in any manner significant to this study However, we cannot assume that they exactly resemble the mean and standard deviation of his whole batch This fact does not affect the differences between marking and re marking which are the prime consideration of this study

The signs attached to the row of Differences between the means are arbitrary When the second mean is higher, we have attached a '+ —

when the mean for re-examining is lower, we have attached a "—" This seems logical (or at least psychological!)

The same comments apply to the SDs (standard deviations) A "+" means the second examiner widened out the distribution, a "—" means that he attenuated, narrowed, or compressed it

Note, however, that in a sense the "direction" is really irrelevant "First" and "second" examiners could easily be interchanged There are some consistent features of the "second" examiners—on the average they have marked a little lower, and with less J-effects However, the far more important factor is the actual differences, regardless of "direction".

Differences between means range up to 3 78, (which is 11% of marks) Differences in SD range up to 1 04 (or 3% of marks)

For the differences between means, * indicates that the difference is significant at the 5% level of confidence, ** that it is significant at the 1% level of confidence, and *** that the difference is significant at the 0 1% level of confidence

The interpretation for the asterisks attached to the SD is the same, except that, unfortunately, we had no tables for significance at the 0 1% level of confidence for this statistic

We may summarize these data as follows, giving the number of differences which are significant at each level of confidence

	001	01	05	NS
M	9	1	4	6
SD	××	3	4	13

	Re-Examiner	
	Same	Diff
Neither M nor SD differences are significant	3	1
Mean difference only is significant	6	3
SD difference only is significant		2
Both M and SD differences are significant	1	4

Remember that if the examiners are *reliable* (in the broad sense of the term) then the differences should be *not* significant We find four cases (3 of self re marking, only one with a different examiner) where both of the differences were small enough to be attributed only to random chance Note, however, that this alone does not guarantee that both sets of marks will be the same The coefficient of correlation must also be taken into account Examiner G, for example, had no significant difference in mean or SD between his two markings But the correlation between the two markings was only + 76, producing an average difference of 2 00 between individual marks

Sixteen out of the 20 pairs were significantly different in mean, SD, or both, indicating that factors other than sheer chance were operating

4 Distribution of Divisions

Those who do not understand means and standard deviations will get a clear picture of what is happening when they look at the next few rows These show how many candidates were awarded each Division, by each examiner This is a good way of knowing the subjective standard that the examiner has in mind when he marks the answer books

In each major column, you will find two columns of figures in this section These show the number (out of 50) awarded each Division the first and second time, i e by the Original and Re-examiner

Let us first examine the cases where an examiner re-examines his own answer books Look at the seventh column, with D D at the top Examiner D awarded two Seconds the first time, none the second, 31 Thirds the first time, only four the second The first time, Examiner D failed only 17 students—the second time he failed 46 Thus the second time his mean was lower and his SD was narrower Except for Examiners C and G, *there was a wide difference in the number awarded each Division by the same examiner, the first and second time he marked his answer books*

The differences were even wider when it was a different examiner who re-examined the answer books For example, look at the fourth column, Examiner A re marking the answer books of Examiner B Where B awarded 11 Second Divisions A has awarded only one Where B failed 10, A has failed 37, in the same batch This is why B s mean is so much higher than A's In the second column, we see what happens when B re-marks A's answer books B awards two Firsts where A awarded none, and nearly three times as many Seconds But the number of Fails remains about the same This explains why B's SD was much larger than A's In both instances B marks much more leniently than A does

5 Pass Percentage

A popular (though very crude and unreliable) measure of ' quality ' is the Pass Percentage A glance along the next three rows shows how little trust can really be placed in this figure The first of these three rows shows the percentage passed in the Original marking, the second row the percentage passed in the Re marking and the third row the difference between these two percentages

Notice that even the same examiner, re examining his own answer books, can differ from himself by as much as 58% in Pass Percentage

(Examiner D) Only two examiners' Pass Percentages (Examiners C and G) remained exactly the same, and Examiner E's was quite close But the remaining seven examiners showed wide changes

Examiners B and I came fairly close to agreeing with A and X (whose answer books they re-marked) in Pass Percentage In examiner B's case this was sheer good luck—means, SDs and percentages in each Division show that he agreed with A on nothing else When Examiner A re marked the answer books of B, the difference in Pass Percentage was 54%

6 Examiner reliability

There is a wide variation among the correlations between each pair of marks

The self correlations of examiners re examining their own answer books ranged from + 61 (Examiner B) to + 94 (Examiner E) The median was + 81, which shows a fair degree of self agreement This compares favourably with such studies done in other countries

When Hindi examiners marked each other's answer books, the product moment coefficient of correlation ranged from + 66 (Examiners A–B), to + 88 (X–E), with a median of + 78 The range (strangely) is not so wide as that of the self-correlations, but the median is nearly the same

The various statistics do suggest that B may have been unusually careless in his second marking If this is so, this is the *only* clear case of this among the 44 examiners in four subjects in this study

But even a relatively high correlation of + 80 (B–A) does not look so good when you examine the differences in means (3 8), SDs (56), and Pass Percentages (80% vs 26%) This clearly illustrates the fact that the correlation coefficient alone is misleading, as it takes no account of *differences in M and SD*

7 Content reliability

The calculations for content reliability (see Chapter VII) were done on the basis of the original marks awarded each batch of answer books (Note that the batches for A, B, C, D are 100 each, for E, F, G, H, I, J are 50 each, and for Examiner X the number of answer books is 300)

The fact that there is a wide variety of content reliabilities for batches of answer books from the *same* examination needs some explanation In the first place, some of these differences may be due to different choices of questions However, probably more important is each examiner's style or method of marking Thus, two examiners may expect two different

kinds of answers (not completely different, of course, but different in significant aspects) from the same question For example, *fresh and* original ideas may be highly prized by Examiner A as a sign of intelligence and creativity, but strongly disliked by traditionalist B who thinks they show the student has not really understood the course Thus, the candidate who happens to answer in a way which would have pleased Examiner A gets a poor mark because, by chance, his examiner is not A but B In a sense, therefore, we do not really have a single examination—*we have as many examination papers as we have examiners* So students sitting for the same examination are really sitting for different examinations and, because the student knows nothing about his examiner, *he does not know which examination he is sitting for* This is in line with Taylor's remark (which repeats in India similar studies done in the West), that the candidate's mark depends more on the luck of who his examiner is than on his actual merit

Another *factor may be the particular choice of options in a particular* examiner's batch of answer books Harper's formula considers *all options to be equivalent*—but if they are not, this could lead to differences in the content reliability In general, however, the options chosen were the same in every batch of answer books

Content reliabilities range from 45 (for B) to 77 (for X) But even this highest value would be considered grossly inadequate in an objective examination

8 Marks reliability

As has been explained earlier, (Chapter VII) the only fair comparison between traditional and objective examinations is on the basis of what we have termed "*marks reliability*" This is the product of examiner and content reliability We have calculated *this only in the cases* where a different examiner has re marked the answer books 'Marks reliability' attempts to estimate the correlation between the marks of a batch of candidates who have answered examination I marked by Examiner P, and examination II marked by Examiner Q

Marks reliabilities range from a low of 36 to a high of 68 Even this highest reliability is not high enough to estimate, accurately, the relative accomplishment of a *class or group* in two different subjects—let alone estimate the true merit of any individual

Of course, the reliability of the Hindi mark based on *three* papers is somewhat higher than the reliability of a single paper However, the difference is not enough to justify the heavy reliance placed on Hindi marks

In a sentence *Hindi marks are unreliable*

9. Standard error of measurement (SEM)

The SEM ranges from 2.64 to 1.89 marks out of 34. (To compare this with other subjects, the range is 3.88–2.78 marks out of 50.) This means that even for the most accurate batch, there are only two chances in three that a given candidate's mark will be within 2 6 out of 34 (3.9 out of 50) of his "true" mark. For about 5 candidates in every 100, the error will be more than 11% of marks, even for this best batch. But, as has been pointed out, our concern should be for the largest (not the smallest) errors that can occur. A SEM of 2.64 means that for one candidate in each hundred (or 1,000 in every lakh), the error will be 20% of marks *or more*. And whether the mark awarded will be 20% *above* or *below* the "true" mark is a matter of sheer chance.

But note that these estimates of errors take into account only the correlation—not the differences in M and SD. Thus they would be true only if the marks of all examiners were statistically scaled to a common standard. More exactly, each SEM shows the range of error which would exist if the re-marking by the second examiner were scaled to the mean and standard deviation of the original marking of the original examiner. The actual range of error, under current unscaled examination marking methods, would probably be much greater (see discussion in Chapter VII).

10. Differences

The last section of Table 12.2 presents the frequency with which each difference between examiners of a single answer book appeared. Note that this does *not* take into account the content reliability (as the SEM does), but only the actual difference between two markers of the same answer book. If each candidate had answered two parallel examinations, marked by different examiners, these differences would be much larger.

The largest difference for the same examiner re-examining 50 of his own answer books was 8 marks. (This is equivalent to 12 out of 50, the same as biology and slightly higher than history.) There were two such differences. The details of one of them (Examiner B) are as follows:

Question	First Marking	Second Marking
1	6	3
2	1½	½
3	2½	2½
4	3	2½
5	3	1
6	3	1½
	19	11

EXAMINERS IN BATCHES OF 50 ANSWER BOOKS EACH

Y	H	X	I	X	J	X
G	H	H	I	I	J	J
Diff	S	Diff	S	Diff	S	Diff
11 86	9 46	11 84	12.22	11 84	11 42	11 84
11 98	9 56	11 56	11 40	11 84	10 48	13.24
+ 12	+ 10	—.28	— 82***	0	— 94***	+1 40***
345	210	259	231	.253	.258	386
.35	48	1 08	3 55	0	3 64	3 63
3 795	2 685	3 775	2.900	3 630	2.736	4 032
2 753	2 743	3.232	2 653	3 540	2.385	3.314
—1 042***	+ 059	— 543*	— 247	— 089	—.350	—.718
1 1		1		1		2
0 4	1	10 9	8 6	9 11	4 3	9 13
1 30	28 22	21 25	39 30	27 27	32 25	20 27
8 15	22 27	18 16	3 14	13 12	14 22	19 10
64%	56%	64%	94%	74%	72%	62%
70%	46%	68%	72%	76%	56%	80%
6%	10%	4%	22%	2%	16%	18%
767	850	874	831	876	726	741
771	521	771	594	771	519	771
591		674		675		.571
2 43		2 16		2 07		2 64
3.57		3 18		3 04		3.88
						1
		1				2
						3
1					2	1
3						
3		2	3	2	3	2
6	5	2	8	8	7	10
18	11	13	9	8	13	10
13	22	21	11	24	16	13
6	12	11	19	8	9	8
2 00	1 18	1.36	1.30	1.44	1 70	2.32

Notice that this candidate is dropped from the *highest* mark in the *Second* Division (19) to the *lowest* mark in the *Third* (11) You will remember from Table 4 3 and Figure 4 1 that several Hindi candidates were changed *two* Divisions from first to second marking, although the change in number of marks was not as much as above (A change from Fail to II requires only a change of 5 marks from 10 to 15)

The largest difference for a pair of examiners was between Examiners D and C, and was 9 marks out of 34 (equivalent to 13 marks out of 50) The details are as follows

Question	First Marking	Second Marking
1	2	0
2	1½	0
3	1½	½
4	2½	½
5	½	½
6	2½	½
	10½	2
	=11	

Since this largest difference occurred once in 500 answer books, the probability is that there would be *errors* ' of 9 or more for at least 200 out of every lakh of candidates Slightly smaller errors would be very much more common (see Table 4 1)

Note that by chance the two examples chosen are both examples of the second examiner lowering the marks In history it was the other way around Changes in both directions are nearly equally common

It is also instructive to look at a case, chosen nearly at random of agreement between two examiners

Question	First Marking	Second Marking
1	3	6
2	1½	0
3	1	½
4	1	0
5	1	1
6	0	0
	7½	7½
	=8	=8

We will let the reader draw his own conclusions from these Marks

The average differences are reported at the bottom of each column in Table 12 2 The smallest is 1 18 (Examiners C and H) and the largest

3 78 (Examiner A re-marking B) Thus the average differences range from 3 5% to 11% of marks But talking of averages reminds us of the student who was refused a freeship because the "average income of your neighbourhood is over five lakhs". Well, there were two houses on the street One was owned by a millionaire, the other by a man who earned only Rs 500 per year But their *average* income of Rs 5,00,250 debarred both their sons from freeships in the university

So it is the largest differences, not the averages, which should concern us Even the Divisions of 36% of the candidates were different in the two markings There is little comfort in the "average", when the examination can be so very unjust to so many

11 Scatter charts

Table 12 3 compares the work of Examiners A and B on their own and on each other's answer books It also compares the re-marking of Examiners C and D with each other

These and Table 12 4 are the individual scatter charts which are summarized in Table 4 3

For example, look at the first chart in Table 12 4, showing what happened when Examiner A re marked his own 50 answer books Out of 26 that he awarded a Third Division the first time (line 3 and column T, which means Total), the second time he reduced 1 of them to a Fail and raised 4 to a Second Conversely, reading downward, of the 11 who were given a Second Division the second time, only 4 had originally been Seconds One had been a First Divisioner, 4 Third Divisioners, and 2 had failed the first time A marked these answer books

The little dashes (—) indicate zero cases at the points which would show perfect agreement on Division They are to help the eye to find the diagonal of perfect agreement in each chart

At the bottom of the chart we also find the words, "Agree—64%" This refers to the fact that 64% of the candidates received the same Class or Division in both markings by Examiner A

The remaining charts may be read in the same way Note that the top row of four charts in Table 12 3 are the self-agreement of the four examiners, the bottom row are their agreements with each other

The left half of Table 12 4 shows the self-correlation of Examiners E, F, G, H, I, J Note that even Examiner E, who had the highest self-agreement (+ 936) spread both his I and II candidates over three Divisions, the second time he marked them

Percentage agreement and *r* are, of course, different measures of agreement One reason why they are not closely related is that a perfect correlation of marks can exist, even if the Divisions are entirely diffe-

TABLE 12.3

PAIRS OF EXAMINERS MARKING THEIR OWN AND EACH OTHER S CANDIDATES IN *Hindi*

Second Examiner—*A*

First Examiner—*A*	F	3	2	1	D	T
D				1	—	1
1			2	4	1	7
2		1	21	4		26
3		7	7	2		16
F						
T	8	30	11	1		50

Agree = 64%
r = 77

Second Examiner—*B*

First Examiner—*B*	F	3	2	1	D	T
D						
1			2	8	2	12
2			14	14		28
3			8	2		10
F						
T	24	24	2			50

Agree = 48%
r = 61

Second Examiner—*C*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1				—		4
2			1	3		24
3		4	20			22
F		18	4			
T	22	25	3			50

Agree = 82%
r = 88

Second Examiner—*D*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1				—		2
2		1	1	—		31
3		28	3			17
F		17				
T	46	4				50

Agree = 40%
r = 78

Second Examiner—*B*

First Examiner—*A*	F	3	2	1	D	T
D					—	
1				—		6
2			2	3	1	35
3		4	16	14	1	9
F		4	5			
T	8	23	17	2		50

Agree = 46%
r = 66

AA—AB = Z = 23
BB—AB = Z = — 06

Second Examiner—*A*

First Examiner—*B*	F	3	2	1	D	T
D					—	
1				—		11
2			2	8	1	29
3			25	4		10
F			10			
T	37	12	1			50

Agree = 30%
r = 80

AA—BA = Z = 07
BB—BA = Z = — 38

Second Examiner—*D*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1				—		4
2			3	1		20
3		6	13	1		26
F		24	2			
T	30	18	2			50

Agree = 76%
r = 82

CC—CD = Z = 20
DD—CD = Z = — 12

Second Examiner—*C*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1				—		3
2			3	—		30
3		19	11			17
F		17				
T	36	14				50

Agree = 56%
r = 76

CC—DC = Z = 37
DD—DC = Z = 05

* difference of 40 is significant at 5% level of confidence
** difference of 53 is significant at 1% level of confidence

rent Another reason is that the Division categories are of unequal size "Fail" covers 11 marks (including 0), "III ' only 4 marks

The right half of Table 12 4 shows what happened to the six equivalent batches into which Examiner X's 300 answer books were divided If you look at the Totals in the extreme right column, of each chart, you will see that the various batches did differ slightly (If there were seven First Divisioners, two of them had to be assigned to one examiner, and only one each to the others Also, some adjustments were made to make the means and SDs equal) However, look at the Total row at the bottom of each scatter chart, comparing all six of them with each other (They are lined up to make this comparison easy) Examiner E awarded 14 Second Divisions, Examiner F only four Note in the left hand column that Examiner E is equally generous and F equally stiff with their own answer books, both the first and the second time Examiner J failed only 10 while Examiner F failed 25 in their virtually identical batches of answer books "Pass ' and "Division' depend as much or more on the examiner as on the candidate's merit

Comparing the two halves of Table 12 4, it is interesting to note that four examiners, G, H, I, and J, have slightly higher correlations with Examiner X than with themselves This is also true of four of the eight comparisons in Table 12 3 To find out if these differences were significant, all r's were translated into Fisher's z, and the differences calculated The last two lines of Table 12 3, and the right hand column of Table 12 4 give the results None of these differences, positive or negative, is significant* All of the Hindi examiners agree equally well whether marking their own or someone else s answer books (The same is true for biology, whereas for history and geometry there are wide differences) The reader may want to speculate for himself as to the reasons why Hindi examiners do not agree with themselves any better than they agree with each other In history we talked about different ' styles of marking For Hindi it may well be that examiners have the same general expectations, and that the differences are therefore, purely due to random error rather than to any consistent difference in the type of answer expected

Examiner E shows a higher per cent agreement (76% vs 74%) but a lower correlation (88 vs 94) with Examiner X than with himself For Examiners G H, and I the reverse is true All of this indicates that per cent agreement and r are by no means equivalent measures of agreement One reason why they are not closely related is that a perfect correlation of marks can exist even if the Divisions are entirely different Another reason is that the Divisions categories are of unequal size "Fail ' covers a range of 11 marks (including 0) ' III only 4

* See footnote on level of significance in chapter XI section 11

12. Self-consistency of examiners' distributions

We have already discussed the product-moment correlation as a measure of self-consistency. Another measure is the extent to which examiners tend to agree with themselves in M, SD, and percentages of marks in each Division.

We get a more critical picture of this when we examine each pair of examiners. Take, for example, Examiners A and B. Remember the experimental design: The hundred answer books originally marked by Examiner A were split into two equivalent batches of 50 each. Similarly, B's 100 answer books were split into two equivalent batches. Fifty of Examiner A's and 50 of B's answer books were then mixed up in a random order, and sent to Examiner A for re-marking. Similarly, B received a mixed batch of both his own and A's answer books. The examiners were told only that "some of the books" would be answer books originally marked by themselves, and some originally marked by others. In response to a questionnaire, examiners assured us that they were generally not aware which books were which.

Here is the comparison of A and B (from Table 12.2):

	A	B		A	B
Mean	11 8	13 4	SD	2 82	3.79
	8 4	12.2		3 29	2.74

In the first row, the 50 answer books to which A originally awarded mean marks of 11.8 were awarded mean marks of 13.4 by Examiner B. The second row must be read backwards. The batch to which B originally awarded a mean of 12.2, was awarded mean marks of 8.4 by A. Thus for both batches, B's mean is consistently higher than A's. However, no such consistency shows in the SDs.

Here is the same comparison for C and D, who were also a matched pair in the experimental design:

	C	D		C	D
Mean	10.2	9.7	SD	2 86	2 95
	8 6	10.7		3 00	2.73

Again, as we also saw in the scatter charts, there is a lack of consistency.

For more detail, let us compare the six examiners who were matched against nearly identical batches of Examiner X's marking:

	Means when re-marking		SDs when re-marking	
	Self	X	Self	X
F	13.7 (2)	12.3 (2)	3.99 (1)	3 44 (2)
F	10.7 (4)	10 7 (6)	3 03 (2.5)	2 67 (6)
G	14 4 (1)	12.0 (3)	3 03 (2 5)	2 75 (5)
H	9 6 (6)	11 6 (5)	2 74 (4)	3 23 (5)
I	11 4 (3)	11.8 (4)	2 65 (5)	3 54 (1)
J	10.5 (5)	13 2 (1)	2 38 (6)	3 31 (3)

TABLE 12 4

SELF AGREEMENT OF SIX EXAMINERS IN *Hindi*

MARKS AWARDED BY SIX *Hindi* EXAMINERS TO APPROXIMATELY EQUIVALENT SAMPLES OF EXAMINER *X*'S CANDIDATES

Second Examiner—*E*

First Examiner—*E*	F	3	2	1	D	T
D					—	
1			4	2	1	7
2		3	9	1		13
3		19	3			22
F	7	1				8
T	7	23	16	3	1	50

Agree 74%
r = 936

Second Examiner—*F*

First Examiner—*F*	F	3	2	1	D	T
D					—	
1			1	—		1
2		4	—			4
3	8	20	3			31
F	13	1				14
T	21	25	4			50

Agree = 66%
r = 84

Second Examiner—*G*

First Examiner—*G*	F	3	2	1	D	T
D					—	
1		1	1	2		4
2		3	19	1		23
3	3	10	5			18
F	2	3				5
T	5	17	25	3		50

Agree = 66%
r = 76

Second Examiner—*H*

First Examiner—*H*	F	3	2	1	D	T
D					—	
1				—		
2			—			
3	7	20	1			28
F	20	2				22
T	27	22	1			50

Agree = 80%
r = 85

Second Examiner—*I*

First Examiner—*I*	F	3	2	1	D	T
D					—	
1				—		
2		2	6			8
3	11	28				39
F	3					3
T	14	30	6			50

Agree = 74%
r = 83

Second Examiner—*J*

First Examiner—*J*	F	3	2	1	D	T
D					—	
1				—		
2		2	2			4
3	11	20	1			32
F	11	3				14
T	22	25	3			50

Agree = 66%
r = 73

Second Examiner—*E*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1				1	—	1
2			2	7		9
3			20	6		26
F	11	3				14
T	11	25	14			50

Agree = 76%
r = 88

Fisher's z Diff between pair: 32

Second Examiner—*F*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1			1	—		1
2			8	2		10
3		9	12	2		23
F	16					16
T	25	21	4			50

Agree = 60%
r = 77

Fisher's z Diff between pair: 19

Second Examiner—*G*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1				1		1
2			8	2		10
3		4	15	2		21
F	11	7				18
T	15	30	4	1		50

Agree = 58%
r = 77

Fisher's z Diff between pair: 00

Second Examiner—*H*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1				1	—	1
2			3	7		10
3		3	17	1		21
F	13	5				18
T	16	25	9			50

Agree = 74%
r = 87

Fisher's z Diff between pair: − 09

Second Examiner—*I*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1				1	—	1
2			4	5		9
3		2	20	5		27
F		10	3			13
T	12	27	11			50

Agree = 70%
r = 83

Fisher's z Diff between pair: − 17

Second Examiner—*J*

First Examiner—*X*	F	3	2	1	D	T
D					—	
1			2	—		2
2			1	8		9
3		1	17	2		20
F	1	9	9	1		19
T	10	27	13			50

Agree = 68%
r = 74

Fisher's z Diff between pair: − 03

* A difference of 40 is significant at 5% level of confidence
** A difference of 53 is significant at 1% level of confidence

The rank orders are given in parentheses beside each number Notice the almost complete lack of consistency The rank order correlations are + 26 for the means and — 27 for the SDs—neither of which is significantly different from zero

In history we found consistent differences in 'standard' of marking between various examiners In Hindi, apparently, the same examiner may be lenient one time and stiff the next, or vice versa The differences in the two markings seem to be due more to the operation of random chance than to any consistent characteristics of individual examiners

If this is so, "scaling" would be even more effective in Hindi than in history in wiping out some of the inequities of the examination At the very least, *it would be wise for the authorities to specify more exactly the distributions which they expect, for "average" or "typical' groups at least* (This is not a radical suggestion It is, in fact, common practice in many countries)

13 Item analysis

One of the most revealing procedures that any examiner can carry out is an item analysis At the simplest level, it takes very little work Merely select a random group of answer books (say 100), arrange in order of merit, divide them into several equal groups and then tabulate how well each of these groups did on each question

We divided our thousand Hindi answer books into five groups of 200 answer books each, on the basis of the first marking Table 12 5 presents the results

We had agreed to keep secret the identity of the Board supplying these data Unfortunately, this precludes our publishing the Hindi question paper However, its organization can be revealed, as it is fairly typical

There were 6 questions, all compulsory, but options allowed within each question as follows

Q 1	Answer 1 out of 11, each of which had 2 alternatives	10 marks
Q 2	Answer 1 out of 9	5 "
Q 3	Answer 1 out of 4	5 "
Q 4	Answer 1 out of 13	5 "
Q 5	Two compulsory parts, the second of which has a choice of 1 from 2 alternatives	2+2 = 4 "
Q 6	Compulsory (no alternatives)	5 "
	Maximum Marks	34

Table 12 5 gives the number of candidates of each ability level answering each question and their marks on that question For example, all 200 in the upper group answered Question 1, and their average marks were 4 92 The average marks of the top group are given at the extreme right, as 16 03

Only one question, Question 4, shows serious non-linearity Compare the mean marks on this question for the third and fourth groups, with their total means

Question 4	Total
1 79	11 26
1 86	10 00

In other words, poorer students did better on this question than better students

Note, also, some inconsistencies in the number of students not answering Question 4 in the various ability groups

The questions differ in difficulty However, in this case this is not "bad' When all are compulsory, it does not matter if one differs from another—for the differences affect all candidates equally

TABLE 12 5

ITEM ANALYSIS OF *Hindi* HIGHER SECONDARY EXAMINATION (FOR DESCRIPTION OF EXAMINATION SEE TEXT)

Question Number Max Marks	1 10	2 5	3 5	4 5	5 4	6 5	Mean Marks
Mean marks of group on question	4 92	2 14	2 37	2 68	1 84	2.18	16 03
(Number in group choosing question)	(200)	(200)	(200)	(197)	(194)	(199)	(200)
Mean marks of group on question	4 11	1 65	1 98	2.16	1.24	1 87	12.98
(Number in group choosing question)	(200)	(199)	(200)	(200)	(195)	(200)	(200)
Mean marks of group on question	3 97	1 47	1 66	1 79	82	1 66	11.26
(Number in group choosing question)	(200)	(199)	(200)	(193)	(195)	(199)	(200)
Mean marks of group on question	3.27	1 33	1.50	1.86	73	1 55	10 00
(Number in group choosing question)	(200)	(191)	(199)	(196)	(190)	(199)	(200)
Mean marks of group on question	1 82	88	1 12	1 19	.39	97	6 08
(Number in group choosing question)	(200)	(185)	(195)	(176)	(191)	(194)	(200)

We wonder, though, if the reader has noticed a serious flaw in this item analysis? This is based only on the parts—*not on the original separate items* If the reader will look at the outline of the question paper above, he will find that there were actually 41 separate questions in this examination For a true item analysis, we should analyse each of these 41 items separately We should then try to find out whether *within each Question,* the various alternatives have equal value (See mathematics for an example of this) As noted earlier, item analysis was a by product, not a central concern of this study To do a detailed item analysis for Hindi would have involved a great deal of work checking which item each candidate had answered, as there was no indication of this on the cover of the answer book (from which most data were taken) It also would have involved the use of a much larger number of IBM cards, and a more complex computer design For both practical and financial reasons, this was not done

So perhaps the *major* value of the item analysis presented here is to show that it is *not* particularly valuable to do an analysis by sections only, rather than by the actual items separately chosen

13

Detailed Statistics—Biology

FOR BIOLOGY we used the zoology paper which consisted of eight questions, from which each candidate had to select five. Six questions had two parts each, with varying marks assigned to each part. If the candidate chose the eighth question, he had to answer four out of five parts. Each question carried ten marks, making a maximum of 50 marks for the paper. (For details see "Item Analysis", at the end of the Chapter.)

1. Distribution of marks

Table 13.1 gives the frequency distribution for each batch of 50 answer books, in biology. These are the detailed frequencies on which the totals given in Table 4.5 are based.

Table 13.1-A shows the frequency distributions for the first marking of each batch of answer books. The top row names the original examiner of each batch. It is his marks which are distributed in this part of the Table.

The hundred answer books of Examiner A were split into two equivalent batches of 50 answer books each. One of these batches was to be re-examined by Examiner A himself, the other by Examiner B, as shown in the second row of the Table. The answer books of Examiners B, C, and D were treated in the same way. The reader can see, from this Table, that the batches were as closely matched as was reasonably possible. Some adjustments had to be made to make the means and SDs equal.

The 300 answer books of Examiner X were split into 6 equivalent batches The matching here, also, is as close as is feasible

Table 13 1-B shows the marks awarded each batch of 50 answer books, the second time they were marked Compare, for example, the first pair of columns The two batches of Examiner A's work, which were closely matched in the first part of the Table, are now awarded quite different distributions of marks by Examiners A and B

The lines across the Tables show the divisions between the Classes or Divisions Notice the striking J effect in the first marking (13 1 A) A large number of candidates (70) were awarded 17 marks (passing), but hardly any (8) were awarded a failing 16 In re marking (13 1-B) this J effect is greatly reduced, but by no means disappears Notice that only Examiner C is completely consistent in using the J-effect both the first and second time with all four batches of 50 answer sheets that he marked (Examiners are referred to the footnote near the beginning of the Chapter on Hindi, for a good rationalization against the sort of educational brinkmanship represented by this tendency to raise borderline candidates by one mark)

There is also a pronounced J effect between the First and Second Divisions in the first marking In the second marking, this is reversed, with the extra cases being piled up just below rather than just above the line! (In history and Hindi, the J merely disappeared) The reverse-J noted between Second and Third Divisions in history and Hindi does not show up here Apparently, there is no special reluctance on the part of biology examiners to award a Second Class

Note that only 34 of the 51 points (50+0) in the "scale" are actually used The marks from 0–3 and from 38–50 are wasted—they might just as well not exist One is inclined to think that the bottom marks may occasionally be used, but the top ones we know are certainly wasted

2 Table 13 2

Table 13 2 summarizes most of the statistical tables produced by the IBM "1620" computer The summary means and medians of these statistics have already been reported in Table 10 1

The top row tells who originally examined the batch of 50 Biology answer books which provided the statistics in the column The second row tells who examined the answer books the second time The third row merely notes (for easier reference) whether the answer books were re-examined by the Same examiner, or a Different one from the original

In the remainder of the Table, "O" refers to the Original marking, and "R" to the Re-examining

3. Mean and standard deviation

The wide variety of means and SDs should perhaps not be taken too seriously. Though the samples of the work of each examiner are not believed to be biased in any significant way, they are also not strictly "random". The exigencies of the situation required that we take the first hundred answer books that we could find to represent each examiner (see Chapter VI). There is no reason to believe that they were essentially different from the rest of the work of that examiner in any manner significant to this study. However, we cannot assume that they *exactly* resemble the mean and standard deviation of this whole batch. This fact does not affect the differences between marking and re-marking, which are the prime consideration of this study.

The signs attached to the row of Differences between the means are arbitrary. When the second mean is higher, we have attached a "+"— when the mean for re-examining is lower, we have attached a "—". This seems logical (or at least psychological!).

The same comments apply to the SDs (standard deviations). A "+" means the second examiner widened out the distribution, a "—" means that he attenuated it.

Note, however, that in a sense the "direction" is really irrelevant. "First" and "second" examiners could easily be interchanged. There are some consistent features of the "second" examiners—on the average they have marked a little lower, and with less J-effect. However, the far more important factor is the actual differences, regardless of "direction".

Differences between means range up to 4.64 (which is 9% of marks) which occurs, surprisingly, when the *same* examiner re-marks his own answer books Differences in SD range up to 2.90 (or nearly 6% of marks.)

For the differences between means, * indicates that the difference is significant at the 5% level of confidence; ** that it is significant at the 1% level of confidence; and *** that the difference is significant at the 0 1% level of confidence.

The interpretation for the asterisks attached to the SDs is the same except that, unfortunately, we had no tables for significance at the 0 1% level of confidence for this statistic.

We may summarize these data as follows, giving the number of differences which are significant at each level of confidence:

	.001	.01	.05	NS
M	6	2	2	10
SD	××	2	2	16

	Re-Examiner	
	Same	Diff
Neither M nor SD differences are significant	3	4
Mean difference only is significant	5	4
SD difference only is significant	1	2
Both M and SD differences are significant	1	

Remember that if the examiners are reliable (in the broad sense of the term), then the differences should *not* be significant We find seven cases (3 of self re marking, 4 with a different examiner) where both the differences were small enough to be attributed only to random chance. Note, however, that this alone does not guarantee that both sets of marks will be the same The coefficient of correlation must also be taken into account When examiner B re-marked the answer books of examiner A, for example, they agreed fairly closely on mean and SD—but their correlation was only + 73, producing an average difference of 3 24 between individual marks

Thirteen out of the 20 pairs were significantly different in means, SD, or both, indicating that factors other than sheer chance were operating

4 Distribution of Divisions

Those who do not understand means and standard deviations will get a clear picture of what is happening when they look at the next few rows These show how many candidates were awarded each Division, by each examiner This is a good way of knowing the subjective standard that the examiner has in mind when he marks the answer books

In each major column, you will find two columns of figures in this section These show the number (out of 50) awarded each Division the first and second time, i e by the Original and Re-examiner

Let us first examine the cases where an examiner re-marks his own answer books Look at the seventh column with D D at the top Examiner D awarded 11 First Divisions the first time, and *none* the second time On the other hand, his Third Divisioners increased from 10 to 21, and Fails from 2 to 7 Thus the second time his mean is lower, and his SD is also somewhat narrower Except for Examiners A, H, I, and J, *there was a wide difference in the number awarded each Division by the same examiner, the first and second time he marked his answer books*

The differences were even wider when it was a different examiner who re examined the answer books For example, look at the fourth column, Examiner A re marking the answer books of Examiner B Though they agreed on the number of First and Third Divisioners Examiner B

awarded twice as many Seconds as A, while A Failed more than three times as many candidates as B, in the same batch This is why B's mean is so much lower than A's Oddly enough, when B re marked A's batch (column two), they agreed fairly closely except in the First Division But notice that B tends to have a narrower spread (a smaller SD) than A in both instances

5 Pass Percentage

A popular (though very crude and unreliable) measure of "quality" is the Pass Percentage A glance along the next three rows shows how little trust can really be placed in this figure The first of these three rows shows the percentage passed in the Original marking, the second row the percentage passed in the Re-marking, and the third row the difference between these two percentages

Notice that even the same examiner, re-examining his own answer books, can differ from himself by as much as 18% in Pass Percentage (Examiner C) While this is a much smaller difference than we found in history and Hindi, it is still not negligible No examiners' Pass Percentages remained exactly the same, although A's, B's, and G's came quite close

Examiners X and H, surprisingly enough, agreed perfectly in Pass Percentage—though we note from the rest of the table that they differed significantly in mean marks, and had a rather low *r* But other Pass Percentage differences went as high as 24% (Examiners B and A)—though when it was B re marking A's answer books, instead of vice versa, the difference was only 4% This latter fact again shows the lack of self-consistency of examiners

6 Examiner reliability

There is a wide variation among the coefficients of correlation between each pair of marks

The self-correlations of examiners re-examining their own answer books ranged from + 75 (Examiner D) to + 91 (Examiner J) The median of + 83 shows a fair degree of self-agreement This compares favourably with such studies done in other countries

When biology examiners marked each other's answer books, the product moment coefficient of correlation ranged from + 67 (Examiners D–C) to + 88 (Examiners X–G), with a median of + 80

But even a relatively high correlation of + 89 (Examiner C) does not look so good when you examine the differences in means (2 3), SDs (95), and Pass Percentages (10% vs 28%) This clearly illustrates the

fact that the correlation coefficient alone is misleading as it takes no account of differences in M and SD

7 Content reliability

The calculations for content reliability (see Chapter VII) were done on the basis of the original marks awarded each batch of answer books (Note that the batches for A, B C, D are 100 each, for E, F, G, H, I, J are 50 each, and for Examiner X the number of answer books is 300)

The fact that there is a wide variety of content reliabilities for batches of answer books from the *same* examination needs some explanation In the first place, some of these differences may be due to *different choices of questions* However, probably more important is each examiner s style or method of marking Thus two examiners may expect two different kinds of answers (not completely different, of course, but different in significant aspects) from the same question In a sense, therefore, we do not really have a single examination—*we have as many examination papers as we have examiners* So students sitting for the same examination are really sitting for *different* examinations—and, because the student knows nothing about his examiner, *he does not know which examination he is sitting for* This is in line with Taylor's remark (which repeats in India similar studies done in the West) that the candidate s mark depends more on the luck of who his examiner is than on his actual merit

Content reliabilities range from 41 (for F) to 68 (for J) with a median of 63 But even this highest value of 68 would be considered grossly inadequate in an objective examination

8 Marks reliability

As has been explained earlier, (Chapter VII) the only fair comparison between traditional and objective examinations is on the basis of what we have termed "marks reliability This is the product of examiner and content reliability We have calculated this only in the cases where a different examiner has re-examined the answer books. "Marks reliability ' attempts to estimate the correlation between the marks of a batch of candidates who have answered examination I marked by Examiner P, and examination II marked by Examiner Q

Marks reliabilities range from a low of 42 to a high of 56 with a median of 50 Note that the range is narrower and the extremes are neither as low nor as high as were found in history and Hindi Apparently the best biology examinations are not so good as the best history and

Hindi exams—but the worst are not nearly so bad, either.

But even the highest reliability of .56 is not high enough to estimate, accurately, the relative accomplishment of a *class or group* in two different subjects—let alone estimate the true merit of any individual.

Of course, the reliability of the biology mark based on *two* papers is somewhat higher than the reliability of a single paper. However, the difference is not enough to justify the heavy reliance placed on biology marks

In a sentence: *Biology marks are unreliable.*

9. Standard error of measurement (SEM)

The SEM ranges from 4.1 to 3.0 marks out of 50. This means that even for the most accurate batch, there are only two chances in three that a given candidate's mark will be within 3 marks of his "true" mark. For about 5 candidates in every 100, the error will be more than 12%, for this best batch. But, as has been pointed out, our concern should be for the largest (not the smallest) errors that can occur. An SEM of 4.1 marks means that for one candidate in each hundred (or 1,000 in every lakh), the error will be 21% of marks *or more*. And whether the mark awarded will be 21% *above* or 21% *below* the "true" mark is a matter of sheer chance.

But note that these estimates of errors take into account only the correlation—not the differences in M and SD. Thus they would be true only if the marks of all examiners were statistically scaled to a common standard. More exactly, each SEM shows the range of error which would exist if the re-marking by the second examiner were scaled to the mean and standard deviation of the original marking of the original examiner. The actual range of error, under current unscaled examination marking methods, would probably be much greater (see discussion in Chapter VII).

10. Differences

The last section of Table 13.2 presents the frequency with which each difference between examiners of a single answer book appeared. Note that this does *not* take into account the content reliability (as the SEM does), but only the actual difference between two markers of the same answer book. If each candidate had answered two parallel examinations, marked by different examiners, these differences would be much larger.

The largest difference for the same examiner re-examining 50 of his own answer books was 12 marks. This was a difference in the two markings

N EXAMINERS IN BATCHES OF 50 ANSWER BOOKS EACH

X G	H H	X H	I I	X I	J J	X J
Diff	S	Diff	S	Diff	S	Diff
20 42 20 68 + 26 315 83	19 28 19 58 +.30 385 78	20 42 22 36 +1 94*** 449 4 32	22 20 21 82 − 38 427 .89	20 42 20 06 − 36 .430 84	20 14 19 10 −1.04** .346 3 01	20 44 20 74 + 30 402 .75
4 639 3 834 − 806**	4 252 4 322 + 070	4 504 5 309 + 806	4 233 5 078 + 845*	4 486 5.270 + 784	5 618 5 818 +.200	4 474 5 114 + 650
1 16 14 26 31 7 5	12 13 23 24 15 13	1 5 15 20 27 18 7 7	2 3 2 2 19 17 12 13 26 24 30 22 3 6 6 13		4 3 2 3 13 12 14 15 21 20 28 27 12 15 6 5	
86% 90% 4%	70% 74% 4%	86% 86% 0%	94% 88% 6%	88% 74% 14%	76% 70% 6%	88% 90% 2%
.879 640 563	.799 427	.802 640 513	.805 483	818 640 524	909 .679	832 640 .532
3 07		3 14		3 10		3 05
2 1 1 8 13 14 11	1 2 3 5 8 3 15 13	1 1 1 2 8 8 6 8 13 2	1 2 1 10 9 8 12 7	1 2 6 6 7 7 10 11	1 3 6 9 11 9 11	1 2 3 3 10 17 7 7
1.70	1 98	3 10	2.42	2.36	2.08	2.34

of one answer book by Examiner D The details are as follows

Question	First Marking	Second Marking
1	5½	2
2	6	3½
4	6	4
7	6½	4
8	6	4
	30	17½=18

The same examiner first awarded this candidate a First Division, and then a Third Division!

The largest difference between a pair of examiners was also 12 marks It was between Examiners D and C, as follows

Question	First Marking	Second Marking
	6	5½
3	8	6
6	5½	2½
7	7	6
8	7	2
	33½	22

It is interesting to note that the widest divergence was on Question 8, in which the candidate had to choose four out of five parts The marks awarded by the two examiners were as follows

1½	1½	2	2=7
1	1	0	0=2

This largest difference of 12 marks occurred twice in 1,000 answer books. Thus the probability is that there would be "errors" of 12 or more for at least 200 out of every lakh of candidates Slightly smaller errors would be very much more common (see Table 4 1)

Note that by chance the two examples chosen were both examples of the second examiner lowering the marks In history it was the other way around Changes in both directions are nearly equally common

It is also instructive to look at a case, chosen nearly at random, of "agreement" between two examiners

Question	First Marking	Second Marking
1	4½	3½
2	2	3
3	4½	3½
7	5	4½
8	3½	5
	19½=20	19½=20

We will let the reader draw his own conclusions from these marks.

The average differences are reported at the bottom of each column in Table 13 2. The smallest is 1.70 (Examiners X and G), and the largest is 4.92 (Examiner D re-marking his own answer books) It is amusing (or is it tragic?) that both the largest difference between means (4 64), and the largest mean difference (4.92) were for biologists re-examining their own answer books, not someone else's! Also, the smallest mean difference (1.70) was between two different biologists, rather than one re-marking his own answer books.

The average differences thus range from 3.4% to nearly 10% of marks. This is approximately the same as found for history and Hindi. However, many people have died of heat stroke in cities whose *average* temperature is only 70° F. So it is the largest differences between examiners that are possible, that should concern us. Even the Divisions of 38% of the biology candidates were different in the two markings. There is little comfort in the "average", when the examination can be so very unjust to so many.

11. Scatter charts

Table 13.3 compares the work of Examiners A and B on their own and on each other's answer books. It also compares the re-marking of Examiners C and D with each other.

These and Table 13.4 are the individual scatter charts which are summarised in Table 4.3.

For example, in Table 13.3 look at the third chart in the top row. This shows what happened when Examiner C re-marked his own 50 answer books. The first time he awarded 4 First Divisions (line 1 and column T, which means Total)—but in the second marking, all 4 of them were reduced to Second Class. On the other hand, out of 22 Seconds in the first marking, 2 were raised to First and 6 lowered to Third in the second marking. Conversely, out of 14 who were Failed in the second marking, only 5 had been considered Failures in the first marking of these answer books, and 9 had been Third Divisioners.

The little dashes (—) indicate zero cases at the points which would show perfect agreement on Division They are to help the eye to find the diagonal of perfect agreement in each chart.

At the bottom of the chart we also find the words, "Agree=58%". This refers to the fact that 58% of the candidates received the same Class or Division in both markings by Examiner C

The remaining charts may be read in the same way Note that the top row of four charts in Table 13 3 are the self-agreements of the four examiners; the bottom row are their agreements with each other

TABLE 13 3

PAIRS OF EXAMINERS MARKING THEIR OWN AND EACH OTHER S CANDIDATES IN *Biology*

OWN

Second Examiner—*A*

First Examiner—*A*	F	3	2	1	D	T
D					—	
1			3	4		7
2		3	15	2		20
3	4	8	4			16
F	4	3				7
T	8	14	22	6		50

Agree = 62%
r = 86

Second Examiner—*B*

First Examiner—*B*	F	3	2	1	D	T
D					—	
1				2		2
2		10	14			24
3	1	15	2			18
F	6					6
T	7	25	16	2		50

Agree = 74%
r = 87

Second Examiner—*C*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1			4	—		4
2		6	14	2		22
3	9	10				19
F	5					5
T	14	16	18	2		50

Agree = 58%
r = 89

Second Examiner—*D*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1		3	8	—		11
2	1	12	14			27
3	4	6				10
F	2					2
T	7	21	22			50

Agree = 44%
r = 75

OTHERS

Second Examiner—*B*

First Examiner—*A*	F	3	2	1	D	T
D					—	
1			5	3		8
2		7	11			18
3	2	12	4			18
F	2	4				6
T	4	23	20	3		50

Agree = 56%
r = 73
AA—AB = Z = 37
BB—AB = Z = 40*

Second Examiner—*A*

First Examiner—*B*	F	3	2	1	D	T
D					—	
1			2			2
2	3	7	9	2		21
3	9	12	1			22
F	5					5
T	17	19	12	2		50

Agree = 52%
r = 76
AA—BA = Z = 30
BB—BA = Z = 33

Second Examiner—*D*

First Examiner—*C*	F	3	2	1	D	T
D					—	
1			3	1		4
2		1	16	1		18
3	4	16	4			24
F	3	1				4
T	7	18	23	2		50

Agree = 72%
r = 80
CC—CD = Z = 30
DD—CD = Z = —13

Second Examiner—*C*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1		3	8	1		12
2	1	10	15			26
3		8	2			10
F	2					2
T	3	21	25	1		50

Agree = 52%
r = 67
CC—DC = Z = 59**
DD—DC = Z = 16

* A difference of 40 is significant at 5% level of confidence
** A difference of 53 is significant at 1% level of confidence

It is interesting to note that the self-correlation of Examiner D (.75) is lower than his correlation with Examiner C (.80) though the difference is not statistically significant There is an even wider difference between the percentage agreement on Division (44% vs 72%) These differences may be partly due to characteristics of the batches of answer books as when both of these examiners are working on identical batches the self agreement is higher (This is shown by each column of scatter charts) Note however that percentage agreement is higher with the other than with the self on both occasions when dealing with identical batches (72% vs 58% and 52% vs 44%) Why this should be so we do not know—but it does serve to demonstrate the fact that *r* and per cent agreement on Division are two different statistics

The last two lines of Table 13 3 give the differences in self and other-correlation All *r* s were translated into Fisher s *z* and the differences calculated Examiners A and B correlate significantly* less well with each other than Examiner B correlates with himself in one comparison but not in the other or in comparison with A s self correlation The same is true of Examiners D and C

The left half of Table 13 4 shows the self correlation of Examiners E F G H I J Note that even Examiner J who had the highest self agreement (+ 909) spread both his II and III candidates over three Divisions when he re marked them Percentage agreement and *r* are of course different measures of agreement One reason why they are not closely related is that a perfect correlation of marks can exist even if the Divisions are entirely different Another reason is that the Division categories are of unequal size Fail covers 17 marks (including 0) III only 6 marks

The right half of Table 13 4 shows what happened to the six equivalent batches into which Examiner X's 300 answer books were divided If you look at the Totals (T) in the extreme right column of each chart you will see that the various batches did differ slightly (If there were eight First Divisioners two examiners had to receive two each while the other four had only one each Also some adjustments were made to make the means and SDs equal) However look at the Total row at the bottom of each scatter chart comparing all six of them with each other (They are lined up to make this comparison easy) Examiner G awarded 31 Third Divisions while Examiner E awarded only 11 Examiner J failed 5 while Examiner E failed 18 in their virtually identical batches of answer books

Comparing the two halves of Table 13 4 it is interesting to note that Examiner G has a higher correlation with Examiner X than with

* See footnote on level of significance in Chapter XI Section 11

himself The difference is not statistically significant, but neither are the differences for Examiners E, F, H, I, and J The point is that all six of these examiners (as well as six out of the eight comparisons in Table 13 3) agree equally well whether marking their own or someone else's answer books The reader may want to speculate for himself as to the reasons why biology examiners do not agree with themselves any better than they agree with each other In history we talked about different "styles of marking". For biology it may well be that examiners have the same general expectations, and that the differences are therefore purely due to random error rather than to any consistent differences in the type of answer expected

Examiner C shows a lower per cent agreement but a higher r with himself than with D The same is true for D with himself and with C All of this indicates that per cent agreement and r are by no means equivalent measures of agreement

The interested reader may make other comparisons for himself

12 Self-consistency of examiners' distributions

We have already discussed the product-moment correlation as a measure of self-consistency Another measure is the extent to which examiners tend to agree with themselves in M, SD, and percentages of marks in each Division

We get a more critical picture of this when we examine a matched pair of examiners Take, for example, Examiners A and B Remember the experimental design The hundred answer books originally marked by Examiner A were split into two equivalent batches of 50 each Similarly, B's 100 answer books were split into two equivalent batches Fifty of Examiner A's and 50 of Examiner B's answer books were then mixed up in a random order, and sent to Examiner A for re-marking Similarly, B received a mixed batch of both his own and A's answer books The examiners were told only that "some of the books" would be answer books originally marked by themselves, some originally marked by others In response to a questionnaire, examiners assured us that they were generally not aware which books were which

Here is the comparison of A and B (from Table 13 2)

	A	B		A	B
	22 9	21 7		5 55	4 93
Mean			SD		
	19 4	22 0		5 87	4 87

In the first row, the 50 answer books to which A originally awarded mean marks of 22 9 were awarded mean marks of 21 7 by Examiner B The second row must be read backwards The batch to which B originally

Second Examiner—*E*

First Examiner—*E*	F	3	2	1	D	T	
D					—		
1			1	5		6	Agree
2	1	4	11	4		20	
3	5	7	2			14	r
F	7	2	1			10	
T	13	13	15	9		50	

Second Examiner—*F*

First Examiner—*F*	F	3	2	1	D	T	
D					—		
1				—	*		Agree
2		13	14			27	
3	5	13	3			21	r
F	2					2	
T	7	26	17			50	

Second Examiner—*G*

First Examiner—*G*	F	3	2	1	D	T	
D					—		
1			2	1		3	Agree
2			15	2		17	
3	1	17	12			30	r
F	—						
T	1	17	29	3		50	

Second Examiner—*H*

First Examiner—*H*	F	3	2	1	D	T	
D					—		
1				—			Agree
2		4	8			12	
3	3	15	5			23	r
F	10	5				15	
T	13	24	13			50	

Second Examiner—*I*

First Examiner—*I*	F	3	2	1	D	T	
D					—		
1			1	1		2	Agree
2		6	11	2		19	
3	3	18	5			26	r
F	3					3	
T	6	24	17	3		50	

Second Examiner—*J*

First Examiner—*J*	F	3	2	1	D	T	
D					—		
1			2	2		4	Agree
2		5	7	1		13	
3	4	14	3			21	r
F	11	1				12	
T	15	20	12	3		50	

* A difference of .40 is significant at
** A difference of .53 is significant at

by A. The
B used a
appears in

) a matched

that earn
o, but tha
marks th
D's mean
consistenc
uch consis

who we
king

g

(1)
(5)
(6)
(2)
(3)
(4)

h numbe
ime patte
cy in SD
ually zero

s involve
(3 2) Th
he avera
ore or l
of marks
that th
h exam
should
be mar
mated
, It wo

awarded a mean of 22 0, were awarded mean marks of 19 4 by A The SDs are read in the same way Note that for both batches B used a narrower distribution of marks than A No such consistency appears in the means

Here is the same comparison for C and D, who were also a matched pair in the experimental design

	C	D		C	D
	22 3	22.1		5 01	5 14
Mean			SD		
	22.3	25 8		4.27	4 62

Again, consistency in SD but *not* in mean One hypothesis that came up was that there may be a true consistency in means also, but that it is masked by the tendency to give slightly lower mean marks the second time When we made this correction, we found that D's means were slightly higher than C's on both occasions, but no such consistency appeared in A and B Nor (as we shall see next) is there any such consistency for the remaining examiners

For more detail, let us compare the six examiners who were matched against nearly identical batches of Examiner X's marking

	Means when re marking		SDs when re marking	
	Self	X	*Self*	X
E	21 9 (2)	20 0 (5)	6 86 (1)	7 41 (1)
F	20 6 (4)	19 3 (6)	3 80 (5)	4.26 (5)
G	23 8 (1)	20 7 (2.5)	3 65 (5)	3 83 (6)
H	19 6 (5)	22 4 (1)	4 32 (4)	5 31 (2)
I	21 8 (3)	20 1 (4)	5 08 (3)	5.27 (3)
J	19 1 (6)	20 7 (2 5)	5 82 (2)	5 11 (4)

The rank orders are given in parentheses beside each number Notice that here, as in the comparison of pairs above, the same pattern emerges—no consistency in means, but striking self-consistency in SDs The rank order correlation for means is — 27 (which is essentially zero); for SDs it is + 77

How to account for this? Except for where Examiner D is involved, the differences in means are mostly quite small (see Table 13 2) This suggests that perhaps examiners agree pretty well on what the average standard should be, and that they vary around this average more or less at random However, they do *not* agree on the *distribution* of marks—how many high and low marks should be awarded It is here that there are *consistent differences in "standard" among examiners* Each examiner applies his own concept of how many should fail, how many should be placed in each Division, to each batch of answer books that he marks Apparently the present instructions and "model papers", etc, issued do not serve to control these differences in "standard" adequately *It would*

be wise for the authorities to specify more exactly the distribution which they expect, for "average" or "typical" groups at least. (This is not a radical suggestion. It is, in fact, common practice in many countries.)

13. Item analysis

One of the most revealing procedures that any examiner can carry out is an item analysis. At the simplest level, it takes very little work: Merely select a random group of answer books (say 100), arrange in order of merit, divide them into several equal groups, and then tabulate how well each of these groups did on each question.

We divided our thousand biology answer books into five groups of 200 answer books each, on the basis of the first marking. Table 13.5 presents the results.

We had agreed to keep secret the identity of the Board supplying these data. Unfortunately, this precludes our publishing the biology question paper. However, its organization can be revealed, as it is fairly typical.

There were eight questions, of which the candidate was to choose any five. The questions carried ten marks each, for a maximum of 50. Question 8 had five parts, of which the candidate (if he chose No. 8) was to answer any four. Each of the other questions (except No. 5) had two parts, both compulsory for any candidate who chose the question. The marks for the parts were distributed in various ways: 2+8, 3+7, 5+5, 6+4, 7+3.

Table 13.5 gives the number of candidates at each ability level choosing each question, and their average marks on that question. For example, 138 in the top group chose Question 1, and their average marks were 5.7. The average total marks of the top group are 28.72, as shown at the extreme right.

Note how different the average marks are for various questions, for students of equal ability (as judged by total marks). For example, a student in the high group who chose Question 8 would receive (on the average) 5.2 marks. But if he chose Question 7, he would receive 6.1 marks—a bonus of nearly one mark just for his choice! The differences in other ability groups is even wider.

Gayen (1961, pp. 51–53) has emphasized the fact that a candidate's total marks are determined partly by his choice of questions. This can be illustrated for the bottom-scoring group (bottom row of Table 13.5), comparing the average marks of those who choose the five easiest as against those who choose the five hardest questions.

of each question separately. The results might have been more revealing However, difference in difficulty in parts is not so serious as differences between questions. This is because differences between parts affect all candidates who choose the question approximately equally. It is the differences between questions which lowers the reliability and validity of marks

A more sophisticated method of item analysis might have revealed these and others facts in more precise detail. This was not done as this was not a central concern of this study. However, as a side-light, the results seem sufficiently revealing of one more source of weakness in the traditional type of examination, to justify the effort put into them.

14

Detailed Statistics—Mathematics

WHEN DISCUSSING objective examinations, how often we hear teachers say, "But mathematics examinations *are* objective'

If by this they mean that there is not much room for disagreement in marking, then of course, they are partly right

However, if by this they mean that mathematics examinations are comparable to objective type examinations then they are absolutely wrong

Objective examinations contain a much wider and more accurate sampling of the candidates knowledge and ability than do traditional mathematics examinations For this reason the results are far more reliable Mathematics examinations of the traditional type are unreliable because the sampling of questions is totally inadequate This defect is further enlarged by the fact that choice of questions is allowed, and the marking is on a scale very different from that used in any other subject

All of this—and much more—will become quite clear as we examine the evidence uncovered in this study

To represent mathematics we used the geometry paper This was the recommendation of the Secretary of the Board, who felt that the marking of the geometry paper was less purely mechanical than the marking of the algebra paper (Whether or not this is actually true would be an interesting subject for some future investigator)

The geometry examination consisted of three groups of questions, with choices allowed within each group. The questions in two of the three groups had two parts each. Maximum marks were 50 for this paper. (For details see "Item Analysis", at the end of this section.)

1. Distribution of marks

Table 14.1 gives the frequency distribution for each batch of 50 answer books, in geometry. These are the detailed frequencies on which the totals given in Table 4.5 are based.

Table 14.1-A shows the frequency distributions for the first marking of each batch of answer books. The top row names the original examiner of each batch. It is his marks which are distributed in this part of the Table.

The hundred answer books of Examiner A were split into two equivalent batches of 50 answer books each. One of these batches was to be re-examined by Examiner A himself, the other by Examiner B, as shown in the second row of the Table. The answer books of Examiners B, C, and D were treated in the same way. The reader can see, from this Table, that the batches were as closely matched as was reasonably possible. Some adjustments had to be made to make the means and SDs equal.

The 300 answer books of Examiner X were split into 6 equivalent batches. The matching here, also, is as close as is feasible.

Table 14.1-B shows the marks awarded each batch of 50 answer books, the second time they were marked. Compare, for example, the fourth pair of columns. The two batches of Examiner D's work, which were closely matched in the first part of the Table, are now awarded quite different distributions of marks by Examiners C and D.

The lines across the Tables show the Divisions between the Classes or Divisions. Notice the striking J-effect in the first marking (14.1-A). A large number of candidates (46) were awarded 17 marks (passing), but hardly any (8) were awarded a failing 16. In re-marking (14.1-B) this J-effect is greatly reduced, but by no means disappears. Notice that only four of the ten examiners (A, C, F, J) are completely consistent in using the J-effect both the first and second time with all four batches of 50 answer books that they examined. (Examiners are referred to the footnote near the beginning of the section on Hindi, for a good rationalization *against* the sort of educational brinkmanship represented by this tendency to raise borderline candidates by one mark.)

There is also a tendency to a J between First and Second Divisions, but within the context of the variation in frequencies of various marks, this is not very pronounced in the first marking. In the second marking it disappears altogether. (The same J appeared in history, Hindi, and

TABLE 14 1 B

FREQUENCY ARDED BY *Second Examiner* TO EACH BATCH OF 50 ANSWER BOOKS IN GEOMETRY

Examiner			E	F	G	H	I	J	X						
Batch for re marking by	A	C	E	F	G	H	I	J	E	F	G	H	I	J	Re-marking
50	1											1		1	4
49	—										1	1	2	—	4
48	—								1		1	—	—	—	6
47	—								2		—	1	1	3	9
46	2							1	—		1	1	1	2	9
45	1							—	1	3	—	2	3	1	11
44	3			1				—	2	2	3	2	1	1	14
43	2			1				—	3	3	2	1	—	2	16
42	1			2	3			—	1	—	—	1	—	2	12
41	—			—	1			—	1	—	2	4	1	—	12
40	1			—	2			—	1	—	1	1	3	—	17
39	2			—	—			2	1	2	1	—	—	1	10
38	1			—	1			1	—	—	—	1	1	3	10
37	2	1		—	1	1		—	3	3	3	—	3	1	20
36	—	—		1	1	—		—	1	1	2	—	2	—	10
35	—	—	1	—	1	1		2	1	2	1	—	1	2	16
34	3	—	—	1	—	—	1	1	2	2	1	1	—	—	13
33	1	—	—	1	2	—	—	2	1	1	1	2	1	2	18
32	3	—	—	—	3	—	1	2	2	2	2	2	—	2	17
31	1	1	—	2	3	—	1	2	4	3	1	1	1	4	27
30	—	—	—	—	—	—	—	3	—	1	2	1	2	3	18
29	—	—	—	3	2	—	—	1	2	4	2	1	1	—	22
28	1	1	—	3	2	1		—	1	1	2	3	—	2	24
27	1	—	—	1	1	—	—	—	3	1	1	—	2	2	17
26	1	1	1	1	3	1	3	—	1	—	—	3	3	—	20
25	1	1	—	3	2	1	2	—	—	1	—	1	2	1	18
24	—	—	1	2	—	2	4	—	—	—	3	2	4	2	32
23	—	—	—	1	2	1	2	1	2	—	—	1	1	—	19
22	2	1	2	1	1	—	3	1	—	4	3	—	2	1	26
21	1	3	—	5	2	1	2	—	—	2	2	2	—	2	26
20	1	3	—	—	1	—	1	1	1	1	—	2	4	4	25
19	2	—	1	2	—	—	2	2	1	2	2	3	1	2	29
18	1	2	1	1	1	1	2	2	2	—	2	—	1	—	23
17	1	2	—	4	1	2	1	3	3	2	—	1	—	—	35
16	—	—	1	—	3	1	1	—	—	—	1	—	—	—	14
15	1	1	—	3	—	2	3	3	1	2	1	—	—	1	27
14	1	3	1	1	1	3	2	2	1	1	—	—	—	—	24
13	2	—	1	1	1	2	2	—	—	—	1	—	1	—	21
12	2	3	1	—	2	1	2	2	1	1	1	2	—	—	37
11	1	1	1	2	2	—	1	—	—	—	—	—	—	—	21
10	—	1	1	2	—	1	—	2	—	—	—	1	—	1	17
9	—	3	1	1	1	2	1	2	—	—	—	1	2	1	23
8	1	3	1	3	—	1	2	1	—	—	—	1	1	1	23
7	1	—	3	—	—	2	2	2	2	1	1	—	—	—	26
6	1	—	4	—	2	2	1	2	—	1	2	1	1	1	18
5	2	3	—	—	—	2	1	1	1	—	—	—	—	—	18
4	—	—	5	—	2	1	1	3	—	1	—	—	—	—	27
3	—	1	7	—	—	4	2	—	—	—	1	1	1	1	24
2	1	2	3	1	—	4	1	2	—	—	—	—	—	—	22
1	1	5	2	—	—	4	2	1	—	—	—	—	—	—	33
0	—	8	5	—	—	4	—	3	1	—	—	—	—	—	36
			6												
Number	50	50	50	50	50	50	50	50	50	50	50	50	50	50	1000

— is used in a coo mark when there were marks awarded above or below th s
Th s is to emparks in certain instances

biology In history and Hindi it disappeared in the second marking, but in biology it was reversed) There is also an interesting reverse-J between Distinction and First Division Apparently mathematics examiners are reluctant to give a Distinction and tend to pile up extra cases at the top mark of the First Division This reverse J also appears in history and Hindi (not in biology)—but is *one Division lower*, i e between First and Second This suggests (as does other evidence) that "*Distinction*" *in mathematics is equivalent to "First" in other subjects* It is quite clearly so psychometrically—the present evidence suggests that it may be psychologically, i e in the minds of the examiners, also Why mathematicians are permitted to use a different scale of educational values is somewhat of a mystery

Another striking difference between the mathematics "scale" and those of other subjects is the use of the entire range of marks, rather than just a portion of it This would generally be considered a desirable feature, if it were not for the fact that the distribution is so extremely flat—indicating (as discussed earlier) that it is a highly artificial scale, not a natural one, as it is unrelated to all we know about human abilities and achievements

The heavy use of zero must also be questioned About three per cent of the candidates are awarded "0 It cannot be too strongly emphasized that, in the context of an examination *zero is not a measurement* It is rather, purely an *indication* of the *inadequacy and invalidity of the examination* (In circumstances where the standard of the test is objectively correct, zeros can indicate defective teaching—but that is obviously not the case here) Suppose we were using a thermometer with no marks below 0 Would the statement "temperature=0°" then have any meaning? No, of course not—because such a thermometer could not distinguish even between temperatures of —1° and —100° A thermometer should show degrees below and above the coldest and hottest situations in which it is expected to be used A valid examination should have a range wide enough so there are no zero and no 100% marks for the group for which it is constructed

Incidentally, there is also interesting *internal* evidence that '0' is not a measurement If it is a valid measurement, how did it happen that there were 27 zeros the first time the answer books were marked—and 36 the second time an increase of 33%?

2 Table 14 2

Table 14 2 summarizes most of the statistical tables produced by the IBM 1620 computer The summary means and medians of these statistics have already been reported in Table 10 1

The top row tells who originally examined the batch of 50 geometry answer books which provided the statistics in the column. The second row tells who examined the answer books the second time. The third row merely notes (for easier reference) whether the answer books were re-examined by the Same examiner, or a Different one from the original.

In the remainder of the Table, "O" refers to the Original marking, and "R" to the Re-examining.

3. Mean and standard deviation

The wide variety of means and SDs should perhaps not be taken too seriously. Though the samples of the work of each examiner are not believed to be biased in any significant way, they are also not strictly "random". The exigencies of the situation required that we take the first hundred answer books that we could find to represent each examiner (see Chapter VI). There is no reason to believe that they were essentially different from the rest of the work of that examiner in any manner significant to this study. However, we cannot assume that they *exactly* resemble the mean and standard deviation of his whole batch This fact does not affect the differences between marking and re-marking, which are the prime consideration of this study.

The signs attached to the row of Differences between the means are arbitrary. When the second mean is higher, we have attached a "+" —when the mean for re-examining is lower, we have attached a "—". This seems logical (or at least psychological!).

The same comments apply to the SDs (standard deviations). A "+" means the second examiner widened out the distribution, a "—" means that he attenuated it.

Note, however, that in a sense the "direction" is really irrelevant. "First" and "Second" examiners could easily be interchanged. There are some consistent features of the "second" examiners—on the average they have marked a little lower, and with less J-effect. However, the far more important factor is the actual differences, regardless of "direction".

Differences between means range up to 2.78 (which is 5.5% of marks). Surprisingly, this largest difference occurs when the *same* examiner re-marks his own answer books. The maximum difference in SD, 4 06 (or 8% of marks) also occurs with the same examiner re-marking his own answer books. The largest mathematics difference in means was smaller than the largest difference for any of the other three subjects—but the largest difference in SDs was second only to the largest biology difference. (However, Table 10.1 showed that on the average geometry and biology examiners did about as well as each other, and only slightly better than history and Hindi, in respect to these differences.)

For the differences between means, * indicates that the difference is significant at the 5% level of confidence, ** that it is significant at the 1% level of confidence, and *** that the difference is significant at the 0 1% level of confidence

The interpretation of the asterisks attached to the SDs is the same, except that, unfortunately, we had no tables for significance at the 0 1% level of confidence for this statistic

We may summarize these data as follows, giving the number of differences which are significant at each level of confidence

	001	01	05	NS
M	9	3	1	7
SD	××	3	3	14

	Re-Examiner	
	Same	Diff
Neither M nor SD differences are significant	4	1
Mean difference only is significant	6	3
SD difference only is significant		2
Both M and SD difference are significant		4

Remember that if the examiners are *reliable* (in the broad sense of the term), then the difference should be *not* significant We find only five cases (4 of self re marking, 1 with a different examiner) where both the differences were small enough to be attributed only to random chance Note, however, that this alone does not guarantee that both sets of marks will be the same The coefficient of correlation must also be taken into account When Examiner C re examined his own answer books for example, he agreed with himself fairly closely in mean and SD—but his self correlation was only 89, producing an average difference of 2 72 between individual marks Two of his marks to the same answer book, in fact, differed by as much as 11 marks out of 50

Fifteen out of the 20 pairs were significantly different in mean SD or both, indicating that factors other than sheer chance were operating On the basis of mean and SD (though not on correlation) the biologists did better than the mathematicians indicating that they had a clearer common standard in mind Which, again shows the "objectivity' of mathematics examinations to be somewhat of a myth

4 Distribution of Divisions

Those who do not understand means and standard deviations will get a clear picture of what is happening when they look at the next few rows These show how many candidates were awarded each Division

by each examiner. This is a good way of knowing the subjective standard that the examiner has in mind when he marks the answer books.

In each major column, you will find two columns of figures in this section. These show the number (out of 50) awarded each Division the first and second time, i.e. by the Original and Re-examiner.

Let us first examine the cases where an examiner re-marks his own answer books. Look at the seventh column, with D D at the top. Examiner D awarded 4 Second Divisions the first time, but raised it to 10 the second time. On the other hand, he reduced his Failures from 35 to 29. Thus the second time his mean is higher, but his SD remains about the same. Only Examiner E is completely consistent in "standard" for both markings. The differences of the others, however, tend to be less than for the other three subjects in this study.

The differences when a different examiner re-examined the answer books averaged about 4 cases (or 8%) in each Division. For example, look at the fourth column, where A re-examined B's answer books They agreed on the number of Distinctions (4) but on nothing else. A awarded more Seconds, but fewer Firsts and Fails—which is why A had a smaller SD than B. When B re-marks A, the same differences appear. B tends to have a lower mean but a wider spread (larger SD) than A in both instances.

5. Pass Percentage

A popular (though very crude and unreliable) measure of "quality" or "standard" is the Pass Percentage. A glance along the next three rows shows how little trust can really be placed in this figure. The first of these three rows shows the percentage passed in the Original marking, the second row the percentage passed in the re-marking, and the third row the difference between these two percentages.

Notice that even the same examiner, re-examining his own answer books, can differ from himself by as much as 12% in Pass Percentage (Examiner D). Of course, this is smaller than the difference found in biology, and much smaller than the differences in Pass Percentage in history and Hindi. Still, it hardly inspires us to place the confidence in Pass Percentage that some people do. Examiner A's Pass Percentage remained the same both times he examined his answer books—though he differed from himself in every other Division. Examiner E's Pass Percentage also remained unchanged, but there must have been differences among the various levels of Failure to produce the highly significant difference in mean shown.

Examiners X and J, surprisingly enough, also agreed perfectly with each other in Pass Percentage—but they differed in number of First

and Second Divisioners and had an average difference of 2 5 in pairs of marks But their close agreement in means and SD seems to ind cate a fairly good agreement in general standard which could have been made much closer by adequate scaling The largest difference in Pass Percentages for d fferent examiners was 10% (while the largest for the same re examining his own papers was 12%)

Is it fair to condemn a teacher because the Pass Percentage in his class is 10% lower than in another teacher s class ? Or should we not rather just attribute this to the general unreliability of traditional examination ?

6 Examiner reliability

There is a wide variation among the coefficients of correlation between each pair of marks *

The *self correlations of examiners re-examin ng the r own books* ranged from 809 (Examiner D) to 988 (Examiner A) The med an of 98 shows a high degree of self agreement This is in fact somewhat *higher than the usual reports of studies done in other countries*

When mathematics examiners marked each other s answer books the product moment coeff cient of correlation ranged from 916 (Examiner *D re-marking C s answer books)* to 965 (*Examiner B re-marking A*) with a median of 95 It is interesting to note that the lowest coeff cient was for self marking not for one re marking another s

But even a correlation of 965 (A B)—*wh ch is very h gh indeed*—begins to lose its glamour when we note the difference between means (2 3) SDs (2 18) and average d fference between marks (3 70)—with some d fferences ranging as high as 15 marks out of 50 This clearly illus trates the fact that the correlation coeffic ent alone is mislead ng as it takes no account of differences in mean and SD

7 Content reliabil ty

The calculations for content reliability (see Chapter VII) were done on the basis of the orig nal marks awarded each batch of answer books (Note that the batches for A B C D are 100 each for E F G H I J are 50 each and for Exam ner X we have six batches of 50 each)

* The reader may at first be incl ned to d spute this The largest and smallest correla t ons are 988 and 809 respect vely However remember that the non l near ty of *r is most pronounced* at the upper extremes If we translate all *r* s for all subjects (see *Table in Append x*) *into* the r equ valent *z* s, we find the variat on for geometry is the *largest* (*w th a z d fference of* 1 35) *H story* comes next w th both H nd and b ology showing substantially less var ation

The fact that there is a wide variety of content reliabilities for batches of answer books from the *same* examination needs some explanation. In the first place, some of these differences may be due to different choices of questions. However, probably more important is each examiner's style or method of marking. Thus, two examiners may expect two different kinds of answers (not completely different, of course, but different in significant aspects) from the same question. In a sense, therefore, we do not really have a single examination—*we have as many examination papers as we have examiners.* So students sitting for the same examination are really sitting for different examinations—and, because the student knows nothing about his examiner, *he does not know which examination he is sitting for.* This is in line with Taylor's remark (which repeats in India similar studies done in the West), that the candidate's mark depends more on the luck of who his examiner is than on his actual merit.

This interpretation is supported by comparing Examiner X's six samples with each other and with the work of other examiners. Remember that the 300 answer books of Examiner X were divided into six more or less equivalent batches of 50 each. The content reliabilities for these batches range from .711 to .778. The difference between these figures is obviously small enough to be attributed only to sampling error. However, the content reliability for Examiner X does differ significantly from, say, that of Examiner A (.904). Separate calculations for the separate batches of Examiner X were done only in the geometry paper. But if any reader wishes to do a similar study for the separate batches of the Examiner Xs in history, Hindi, and biology, the authors will be happy to supply him with all the necessary raw data.

Note that content reliabilities range from .711 (for one sample of Examiner X) to .904 (A). Of the five examinations (i.e. the "same" examination marked according to five different systems) for which *content reliabilities were calculated, only A's would be considered really* adequate in an objective examination. These reliabilities are substantially higher than the .63 estimated by Gayen (1961, page 71).

8. Marks reliability

As has been explained earlier (Chapter VII), the only fair comparison between traditional and objective examinations is on the basis of what we have termed "'marks reliability". This is the product of examiner and content reliability. We have calculated this only in the cases where a different examiner has re-marked the answer books. "Marks reliability" attempts to estimate the correlation between the marks of a batch of candidates who have answered examination I marked by Examiner P, and examination II marked by Examiner Q

Marks reliabilities range from a low of 68 (D–C and X–J) to a high of 87 (A–B), with a median of 77. The highest of these might be considered acceptable in an objective examination, but *only two out of the ten coefficients* even begin to approach adequate reliability to differentiate among individuals The best that can be said for them is just that they are better than those of any of the three other subjects in this experiment

Of course, the reliability of the mathematics mark based on *two* papers is somewhat higher than the reliability of a single paper If the correlation between the two papers is quite high, the reliability of the total mark may approach the minimum acceptable However, we do not know the correlation, unfortunately—but it is this author's hunch that it is not likely to be very high Even at best, only some mathematics marks would be adequately reliable for differentiating among individuals

In a sentence *Many mathematics marks, perhaps a majority, are unreliable*

9 Standard error of measurement (SEM)

This unreliability is most clearly brought out by the wide SEM, which reflects not only the total reliability, but also the abnormally wide range (SD) of Mathematics marks The SEM ranges from 6 17 to 3 47 marks out of 50 This means that even for the most accurate batch, there are only two chances in three that a given candidate's mark will be within 3½ marks of his "true' mark For about 5 candidates in every 100, the error will be more than 14%, even for this best batch But, as has been pointed out, our concern should be for the largest (not the smallest) errors that can occur An SEM of 6 2 marks means that for one candidate in each hundred (or 1,000 in every lakh), the error will be 32% of marks *or more* And whether the mark awarded will be 32% *above* or 32% *below* the 'true" mark is a matter of sheer chance

But note that these estimates of errors take into account only the correlation—not the differences in M and SD Thus they would be true only if the marks of all examiners were statistically scaled to a common standard More exactly, each SEM shows the range of error which would exist if the re marking by the second examiner were scaled to the mean and standard deviation of the original marking of the original examiner The actual range of error, under current unscaled examination marking methods, would probably be much greater (see discussion in Chapter VII)

10 Differences

The last section of Table 14 2 presents the frequency with which each difference between examiners of a single answer book appeared Note that this does *not* take into account the content reliability (as the SEM does), but only the actual difference between two markers of the same answer book If each candidate had answered two parallel examinations, marked by different examiners, these differences would be much larger

The largest difference for the *same* examiner re examining 50 of his own answer books was 20 marks (This was even larger than the largest difference for a different examiner, which was 15) This was a difference in the two markings of one answer book by Examiner D It is instructive to list the marks in the parts of each question, as well as the total The details are as follows

Question	First Marking	Second Marking
5	1½+0 = 1½	4+4 = 8
6	0 +3½ = 3½	4+4 = 8
7	0 +0 = 0	0+0 = 0
9	0 +4 = 4	1+4 = 5
12	0 = 0	1 = 8
	9	29

The same examiner first failed this candidate, then raised him to het top of the Second Division ! In another case, this examiner had an 18 mark difference

The reader may be tempted to think that this examiner may just have been marking the answer books "at random", and not taking the experiment seriously Of course, if there was undue carelessness (the hyphen is not a typographical error), there is no way of knowing whether this carelessness was in the original or in the second marking However, if the marking was purely at random, then there would be no consistency whatsoever in the marks As a matter of fact, however, this examiner does tend to give the questions roughly the same rank order each time he marks them—indicating clearly that he did re-read them, and did not just "invent the marks" Furthermore, no "at random" marking would produce a self correlation of 81 So we seem to have to conclude that the examiner has somewhat variable standards Since his errors are not *very* much larger than those of others, we must conclude that there are other "experienced examiners" whose marks are just as untrustworthy

BY TEN EXAMINERS IN BATCHES OF 50 ANSWER BOOKS EACH

X G	H H	X H	I I	X I	J J	X J
Diff	S	Diff	S	Diff	S	Diff
1 70 8 70 2 50*** .491 5.092	10 24 10 56 + 32 .257 1.245	31 20 28 62 —2 58*** 556 4 640	15 30 16 00 + 70** .215 3 271	31 18 28 94 —2 24*** 461 4 859	19 26 19 12 — 14 295 475	31 18 31 06 — 12 461 260
11.212 11 780 + 568	9 456 9 533 + 077	11 137 12 665 +1 527*	8 744 8 777 + 034	10 895 11 860 + 965*	11 798 12 079 + .281	10 895 10 689 — .206
16 12 13 13 9 8 7 9 5 8	2 2 6 6 6 4 36 38	17 16 9 7 12 11 8 8 4 8	2 3 11 11 10 11 27 25	16 13 10 10 12 13 8 8 4 6	3 4 13 12 1 2 13 9 20 23	16 16 10 14 13 7 7 9 4 4
90% 84% 6%	28% 24% 4%	92% 84% 8%	46% 50 4%	92% 88% 4%	60% 54% 6%	92% 92% 0%
956 778 744	982	953 759 723	985	962 743 715	985	955 711 679
5 67		5 86		5 81		6 17
1 1 1 1 2 4 5 5 5 12 9 4	2 2 6 5 14 21	1 1 2 3 3 1 4 5 5 5 12 8	3 4 9 19 15	1 1 2 2 3 2 3 4 5 17 10	3 3 5 8 19 12	1 2 3 2 4 9 9 14 6
3 38	1 20	3 42	1 22	1 64	1 54	2 52

The largest difference between a pair of examiners was 15 marks It was between Examiners A and B, as follows

Question	First Marking	Second Marking
1	4 +1 = 5	1½+0 = 1½
2	2½+1 = 3½	1 +0 = 1
4	3 +2 = 5	0 +1 = 1
7	4 +0 = 4	3½+0 = 3½
9	0 +1 = 1	0 +0 = 0
12	3 = 3	0 = 0
	21½=22	7

This largest difference of 20 marks occurred once in 1,000 answer books Thus the probability is that there would be "errors" of 20 or more for at least 200 out of every lakh of candidates Slightly smaller errors would be very much more common (see Table 4 1)

Note that by chance one of the above examples was an example of the second examining raising the marks, and the other an example of lowering Changes in both directions are nearly equally common

It is also instructive to look at a case, chosen nearly at random, of "agreement" between two examiners

Question	First Marking	Second Marking
1	3½+2 = 5½	4 +3 = 7
2	3 +1 = 4	3 +0 = 3
5	3½+0 = 3½	2½+1 = 3½
7	1 +0 = 1	0 +0 = 0
9	0 = 0	0 = 0
11	0 = 0	0 = 0
	14	13½=14

We will let the reader draw his own conclusions from these marks

The average differences are reported at the bottom of each column in Table 14 2 The smallest is 1.20 (Examiner H) and the largest is 4 06 (Examiner D re examining his own answer books) It is amusing (or is it tragic ?) that both the largest difference between means (2 78) and the largest mean difference (4 06) were for a mathematician re marking his own answer books, not someone else s ! Others did slightly better, of course

The average differences thus range from 2 4% to 8% of marks This is not very different from what was found for history and Hindi

But when speaking of averages we are painfully reminded of the *220-volt* refrigerator which burned out even though the *average* electric current was just 220 volts. So it is the largest difference between examiners that are possible that should concern us. One out of every 17 of the differences in geometry was 8 marks or more out of 50 (16%). Even the Divisions of 23% of the geometry candidates were different in the two markings. There is little comfort in the "average", when the examination can be so very unjust to so many.

11. Reliability and scaling

Perhaps the most striking general conclusion is that mathematics examinations would probably profit more than any of the others by the introduction of scaling. *Any* method of scaling would help. All that is needed is to apply some judgement (or else a statistical formula), rather than using just simple arithmetic in calculating "per cent of marks". The mathematicians, of all people, should not feel themselves bound to such elementary (and, really, rather irrational) procedures.

Geometry examinations show a higher examiner reliability than any other. They also show a higher content and total reliability. The differences between the means are smaller than for any other subject. But all of these advantages are thrown away by using a fantastically large SD, and a fantastically flat distribution curve—a distribution which we know is artificial (or an artifact of the marking method), because it does not fit with any scientific knowledge of the distribution of human abilities. This large SD also produces unnecessarily large differences between marks, and unnecessarily large SEM's—making geometry, in spite of its high correlation coefficients, actually, in the practical sense, *the least reliable of the four papers studied in this report* (see Figure 4.6).

12. Scatter charts

Table 14.3 compares the work of Examiners A and B on their own and on each other's answer books. It also compares the re-marking of Examiners C and D with each other.

These and Table 14.4 are the individual scatter charts which are summarized in Table 4.3.

For example, look at the second chart in the top row of Table 14.3. This shows what happened when Examiner B re-marked his own 50 answer books. The first time he awarded 8 First Divisions (line 1 and column T, which means *total*)—but in the second marking, 2 of these were reduced to Third, 1 reduced to Second, and 1 raised to Distinction. Reading the chart up and down, we find that there were 10 Third

Divisions awarded the second time Out of these 2 had originally been Firsts, I a Second, and 3 had been failed by this same examiner the first time he marked the answer books

At the bottom of the chart we also find the words, "Agree 68%" This refers to the fact that 68% of the candidates received the same Class or Division in both markings by Examiner C.

The little dashes (—) in some of the charts indicate zero cases at the points which would show perfect agreement on Division They are to help the eye to find the diagonal of perfect agreement in each chart.

The remaining charts may be read in the same way Note that the top row of four charts in Table 14 3 are the self agreements of the four examiners, the bottom row are their agreements with each other

Notice the peculiar relationship between A's and B's self-consistency and other consistency When A re marks A he correlates higher with himself (988) than when he re marks B (952) However, B correlates more highly with A in both instances (965, 952) than with himself (912) This, and other comparisons within these three scatter charts suggest strongly that B was more careless in his original (i e the actual Board exam) marking than in the re examination of his answer books

The left half of Table 14 4 shows the self-correlation of Examiners E, F, G, H, I, J Note that even Examiner I, with a 985 self agreement, spread most of his re examined candidates over three divisions Compare, also, the other highest self-correlations—I, J, G, and (the highest of all) A

It is interesting to note Examiner E *This is the ONLY case out of forty in the entire study, where there was perfect self-agreement on Divisions* And yet Examiner E did not have the highest self-*r*, the difference between his means was highly significant, and he averaged 1.28 marks difference between the two markings, with one answer book showing a difference of 5 marks out of 50 (10%) the two times So his perfect Per cent Agreement here seems to be due primarily to the fact that he failed nearly all of his candidates both times

Percentage agreement and *r* are, of course, different measures of agreement One reason why they are not closely related is that a perfect correlation of marks can exist even if the Divisions are entirely different. Another is (as we have seen) that the Division categories are of unequal size "Fail" covers 17 marks (including 0), "III" only 6 "Distinction" is twice as wide as any other category, also And not even Second and Third are equal !

The other half of Table 14 4 shows what happened to the six equivalent batches into which Examiner X's 300 answer books were divided If you look at the Totals in the extreme right column, of each chart, you will see that the various batches did differ slightly This was

forced by the process of dividing Also, some adjustments were made to make the means and SD's equal However, look at the Total row at the bottom of each scatter chart, comparing all six of them with each other (They are lined up to make this comparison easy) The differences are not nearly so great as in other subjects However, Examiner H, for example, awarded 7 First Divisions, while Examiner J awarded 14 (twice as many) to a nearly identical batch Note that Examiner J is equally generous with his own answer books (first half of the Table), while Examiner H is equally stingy (But see below on self-consistency) Distinctions range from 10 to 16 per batch, and Failures from 4 to 8

Do examiners agree with themselves any better than they agree with each other? Comparing the two halves of Table 14 4 gives a partial answer to this question, as did the comparison of the two rows in Table 14 3 Sepecifically, the information is given at the extreme right of Table 14 4, and at the bottom of Table 14 3 All r's were translated into Fisher's z, and the differences calculated Two thirds of the 14 pairs of differences are significant*—and in three of them (as we have seen) r is significantly higher for other- than for self correlation Specifically, five differences are significant and four highly significant (And two more would be significant if we used the one-tailed test)

The reader may wish to speculate as to why self-agreement is so much higher than other-agreement in geometry In biology and Hindi we found no such difference In history, where there was a difference, we talked about "styles of marking Apparently there is no general agreement among geometry examiners (as there is among Hindi and biology examiners) as to what is required Possibly this is because an incorrect answer may be correct in some stages, but examiners may not agree as to how many marks to award for each correct stage Experts in this subject may have other explanations In any case, it seems evident that *more specific instructions to examiners* are probably needed Extensive research in the West indicates that such instructions can greatly increase the reliability of marking

The above facts combined with wide differences in mean, SD, and number of cases in each Division, show that in geometry *standards* of various examiners also differ (More of this later)

It should also be noted (again) that a good deal of the wide scatter in the various charts in these three tables is due *primarily* to the abnormal scale used by mathematicians in marking their answer books If their marks were properly scaled, as is done in many countries, and has been repeatedly recommended in India (e g Mahalanobis, 1934, Taylor, 1963), Percent Agreements on Divisions would rise to close to 100%

* See footnote on level of significance in Chapter XI, section 11

TABLE 143

PAIRS OF EXAMINERS MARKING THEIR OWN AND EACH OTHER S CANDIDATES IN *Mathemat cs*

OWN

Second Exam ner—*A*

First Exam ner—*A*	F	3	2	1	D	T
D				1	13	14
1				8	2	10
2			4			4
3		6	2			8
F	14					14
T	14	6	6	9	15	50

Agree = 90%
r = 988

Second Exam ner—*B*

First Exam ner—*B*	F	3	2	1	D	T
D				1	4	5
1			2	4	1	8
2	1	1	3			5
3	4	4	2			10
F	19	3				22
T	24	10	6	5	5	50

Agree = 68%
r — 912

Second Exam ner—*C*

First Exam ner—*C*	F	3	2	1	D	T
D					—	
1				—	1	1
2		1	—			1
3	4	6	1	1		12
F	35	1				36
T	39	8	1	1	1	50

Agree = 82%
r 894

Second Exam ner—*D*

First Exam ner—*D*	F	3	2	1	D	T
D					—	
1				1		1
2			4			4
3	3	4	3			10
F	26	6	3			35
T	29	10	10	1		50

Agree — 70%
r = 81

OTHERS

Second Exam ner—*B*

First Examiner—*A*	F	3	2	1	D	T
D				2	13	15
1			2	4	1	7
2		2	3			5
3	5	3	1			9
T	13	1				14
T	18	6	6	6	14	50

Agree = 72%
r = 965

AA—AB = Z = 64**
BB—AB = Z = — 48*

Second Exam ner—*A*

First Exam ner—*B*	F	3	2	1	D	T
D				1	3	4
1			5	4	1	10
2		1	4	1		6
3		2	5			7
T	18	5				23
T	18	8	14	6	4	50

Agree = 6?%
r = 952

AA—BA = Z = 82**
BB—BA = Z = — 30

Second Exam ner—*D*

First Exam ner—*C*	F	3	2	1	D	T
D					—	
1				—		3
2			1	2		11
3	5	3	3			36
F	33	3				
T	38	6	4	2		50

Agree = 71%
r = 916

CC—CD = Z = — 14
DD—CD = Z — 43*

Second Examiner—*C*

First Examiner—*D*	F	3	2	1	D	T
D					—	
1				1		1
2	1	1	2	1		5
3	2	6	1			9
F	31	4				35
T	34	11	3	2		50

Agree = 80%
r = 920

CC—DC = Z = — 17
DD—DC — Z = — 46*

* difference of 40 is significant at 5% level of confidence
** difference of 53 is significant at 1% level of confidence

Those who still do not believe that the mathematics scale is different are referred to Harper (1963) where it is shown that students of *equal* ability (according to an outside criterion) receive *different* marks in biology and mathematics In the instance quoted, biology students with X amount of ability received an average mark of 64, whereas mathematics students of the same amount of ability were awarded an 82 At the other end, biology students in the lowest ability group were awarded an average of 28 per cent marks while equivalent mathematics students received only 17 So 64=82, and 28=17 Only in the middle, 45% marks, were the two scales really equivalent

There are many other interesting things in Tables 14 3, and 14 4 Note the wide scatter when Examiner B re examines his own answer books, as well as when Examiner A re marks B s Note how far Examiner C spread out Examiner D s 5 Second Divisioners

The interested reader may make other comparisons for himself

13 Self consistency of examiners' distributions

We have already discussed the product moment correlations as a measure of self consistency Another measure is the extent to which examiners tend to agree with themselves in M, SD, and percentages of marks in each Division

We get a more critical picture of this when we examine matched pairs of examiners Take, for example, Examiners A and B Remember the experimental design The hundred answer books originally marked by Examiner A were split into two equivalent batches of 50 each Similarly, B's 100 answer books were split into two equivalent batches Fifty of Examiner A's and 50 of Examiner B's answer books were then mixed up in a random order, and sent to Examiner A for re marking Similarly, B received a mixed batch of both his own and A's answer books The examiners were told only that "some of the books" would be answer books originally marked by themselves, some originally marked by others In response to a questionnaire examiners assured us that they were generally not aware which books were which

Here is the comparison of A and B (from Table 14 2)

	A	B		A	B
	26 3	24 0		14 47	16 66
Mean			SD		
	21 1	19 7		10 44	11 73

In the first row, the 50 answer books to which A originally awarded mean marks of 26 3 were awarded mean marks of 24 0 by Examiner B The second row must be read backwards The batch to which B originally

awarded a mean of 19.7 were awarded mean marks of 21.1 by A. The SDs are read in the same way. Note that for both batches, *B used a wider* distribution of marks than A, and had a lower mean.

Here is the same comparison for C and D, who were also a matched pair in the experimental design:

	C	D		C	D
	11.7	11.4		6.76	8.63
Mean			SD		
	11.1	11.1		9.50	8.45

Examiners C and D do not seem to show the consistency that A and B did.

For more detail, let us compare the six examiners who were matched against nearly identical batches of Examiner X's markings:

	Means when re-marking		SDs when re-marking	
	Self	X	Self	X
E	7.7 (6)	29.2 (2)	7.79 (6)	12.05 (2)
F	22.1 (2)	28.5 (6)	9.80 (3)	10.83 (5)
G	24.6 (1)	28.7 (4)	10.79 (2)	11.78 (4)
H	10.6 (5)	28.6 (5)	9.53 (4)	12.66 (1)
I	16.0 (4)	28.9 (3)	8.78 (5)	11.86 (3)
J	19.1 (3)	31.1 (1)	12.08 (1)	10.69 (6)

The rank orders are given in parentheses beside each examiner. As in C–D above, there is no consistency. The rank order correlations are actually —.31 and —.77 for means and SDs respectively. If there are any consistencies (and we seemed to discover some in Table 14.4), they are completely masked by the wide variety of means and SDs used. It would seem at least possible that Examiner X's batch just happened to be one of exceptionally high ability (every examiner awarded higher means to X's answer books than to his own), and also an unusually wide range of ability (all but one gave larger SDs). If we also assume an unusual amount of variability in Examiner E to J's own batches, these two facts mask any consistencies that might otherwise appear.

This points up the fact that scaling each examiner's work *separately* on an arbitrary statistical basis (as some advocate) may lead to considerable injustice. There are three possible approaches:

(a) The answer books may be randomized before sending to the examiners, as suggested by Taylor (1963). Then purely statistical scaling can be applied.

(b) Or a more judgemental scaling method may be used, as recommended by Gulliksen (1950) and adapted to Indian conditions by Harper (1963). For large groups, the Board would provide detailed

instructions on "standards" and "distributions" of average or typical groups

(c) If examiners do not wish to exercise their judgement, there is a simple method whereby each examiner's "personal equation" (so to speak) can be determined This is described by Harper (1967, foot note on page 36), and in the Chapter on scaling in this book

14 Item analysis

One of the most revealing procedures that any examiner can carry out is an item analysis At the simplest level, this takes very little work Merely select a random group of answer books (say 100) arrange in order of merit, divide them into several equal groups and then tabulate how well each of these groups did on each question

We divided our thousand geometry answer books into five groups of 200 answer books each, on the basis of the first marking Table 14 5 presents the results

We had agreed to keep secret the identity of the Board supplying these data Unfortunately, this precludes our publishing the mathematics (geometry) question paper However, its organization can be revealed, as it is fairly typical

There were three compulsory groups, with a choice of questions within each group, as follows

Group A Answer 3 out of 6
each had 2 parts, carrying 4 marks each (total $8\times3=24$)

Group B Answer 2 out of 4
each had 2 parts, one carrying 5 marks, the other 4 marks
(total $9\times2=18$)

Group C Answer 1 out of 3, only one part each (total 8)

Thus the candidate had to select a total of 6 questions, for a total of 50 marks*

Table 14 5 gives the number of candidates at each ability level choosing each question, and their average marks on that question

It is unfortunate that Questions 12 and 13 had to be lumped into the same column The reason for this is simple The Board's standard answer book cover provides only for 12 questions (Why are not separate covers printed for separate subjects?) So the examiners had no choice but to put marks for both Questions 12 and 13 in the last space This was not discovered by us until far too late for the very large job of sorting

* Though we were not looking for them errors in marking sometimes turned up For example at least one geometry candidate answered seven questions—instead of the prescribed six—and received credit for all seven Such errors do not occur when all candidates are required to attempt all questions, as is recommended by all experts in educational measurement

through all one thousand answer books, and finding which candidates had answered question 12 and which Question 13.

Look at the last column, marked Total. These are the average marks of each group of 200 candidates. From the averages it is evident that the five groups are *quite steeply graded* in ability level, as reflected in total marks.

Now look down the first column, i.e. the column of mean question marks on Question 1. The top group had an average of 5.5 on Question 1. The second group's average was 4.2, but the third's was 4.9. Even though, on the Total the third group was 11 marks below the second group, it did *better* than the higher group on Question 1. There are several other such reversals: Questions 3, 5, and 6 show them between various groups. These reversals produce negative validity, at least at these points—relatively, they lower the marks of higher students, and raise those of lower candidates, quite irrationally.

Now look along the top row—or any other row. Notice how different the mean marks are for supposedly equivalent questions, answered by students of equal ability (as judged by total marks). These differences must, of course, be examined separately in each Group of questions. Look at Group A (Questions 1–6) for the top group. A student who chooses Question 1 gets (on the average) 5.5 marks, while one who chooses Question 5 gets 7.9 marks—a bonus of 2.4 marks just for his choice! In Group B (Questions 7–10), the marks awarded Question 10 are 4 marks out of 9 (44%) higher than those awarded Question 8, for this ability group. For other ability groups, some differences are larger, some smaller.

Gayen (1961, pp. 51–53) has emphasized the fact that a candidate's total marks are determined partly by his choice of questions This can be illustrated with, say, the second-scoring group (second row) of Table 14.5. Compare the average marks of those who choose the easiest as against those who choose the hardest questions in each group. (In Group C we will have to treat Questions 12 and 13 as one question, because we lack separate data, as explained above.)

Question	Marks	Question	Marks
A 2	7.1	A 3	3 6
5	8 0	4	4.2
6	7 8	1	4.2
B 9	4.1	B 8	2.1
10	5.6	7	4 0
C 12/13	5 4	C 11	4 1
	38 0		22.2

This is nearly 16 marks out of 50—32% of marks—difference *just between different combinations of questions*. Among students of equal ability

ITEM ANALYSIS OF *Mathematics* HIGHER SECONDARY EXAMINATION (FOR DESCRIPTION OF EXAMINATION SEE TEXT)

	Group A						Group B				Group C			
Question Number	1	2	3	4	5	6	7	8	9	10	11	12	13	Mean Marks
Max Marks	8	8	8	8	8	8	9	9	9	9	8	8		
Mean marks of group on question	5 45	7 77	6 62	7 18	7 90	7 10	5 13	3 14	7 09	7 28	6 40	6 99		41 20
(Number in group choosing question)	(89)	(92)	(20)	(63)	(181)	(151)	(92)	(14)	(159)	(132)	(74)	(125)		(200)
Mean marks of group on question	4 23	7 08	3.58	4 15	7 96	7 81	4 01	2 15	4 10	5 61	4 14	5 43		29 50
(Number in group choosing question)	(161)	(83)	(32)	(39)	(167)	(104)	(131)	(47)	(106)	(70)	(55)	(80)		(200)
Mean marks of group on question	4 91	3 76	4 04	4 12	5 60	3 88	2 70	1 76	1 42	1 26	1 02	1 57		18 56
(Number in group choosing question)	(134)	(87)	(25)	(29)	(181)	(136)	(133)	(33)	(113)	(101)	(80)	(103)		(200)
Mean marks of group on question	3 19	2 93	2 32	2 42	4 18	3 83	1 54	83	72	61	0	1 61		12 07
(Number in group choosing question)	(195)	(105)	(37)	(62)	(135)	(59)	(136)	(39)	(103)	(66)	(70)	(58)		(200)
Mean marks of group on question	1 60	53	65	91	93	75	40	25	24	46	0	21		3 50
(Number in group choosing question)	(189)	(78)	(49)	(55)	(141)	(67)	(133)	(59)	(88)	(65)	(75)	(50)		(200)

(i e with the same average total marks) Some could receive a III Division and some a Distinction, based only on a lucky or unlucky choice of questions ! Obviously, a candidate's Division quite often depends on which questions he chooses, even more than on how much he knows If the reader cares to work it out for himself, he will find that in every group candidates of *equal* ability can get *different* marks just by happening to choose one or another set of questions We have argued elsewhere that the examiner who allows a choice of questions is stating that he believes the questions to be of equal difficulty Since this is very obviously *not* true, and since the candidate has no way of knowing which questions the examiners will mark ' stiffly ' and which "easily", the traditional essay type examination ends up as a sort of guessing game The student who is lucky enough to make the right guesses gets higher marks than he deserves—the poor fellow who guesses wrong suffers unjustly

There is universal agreement among experts in the science of educational measurement that (except where necessitated by the use of different courses or texts) there is no justification for allowing alternatives, or a choice of questions in any examination That this choice factor is a major source of unreliability and, more important, of invalidity has been demonstrated by Harper (1962) The ' chance ' factor in this type of examination can only be described as huge

Another indication of inconsistency is the fact that Question 2 is easier than 1 for the highest group—but for the lowest group it is the other way around, with Question 1 receiving the higher marks This implies very different distributions of marks (wide and narrow) for various questions The scaling of marks for individual questions might help a great deal in increasing validity of total marks—although of course, eliminating options is both easier and more effective

It is unfortunate that circumstances did not permit a more detailed tabulation of each *part* of each question separately The results might have been even more revealing However, difference in difficulty in parts is not so serious as differences between questions This is because differences between parts affect all candidates who choose the question approximately equally It is the differences between questions which lowers the reliability and validity of marks

A more sophisticated method of item analysis might have revealed these and other facts in more precise detail (see Gayen 1961) This was not done as this was not a central concern of this study However, as a side-light, the results seem sufficiently revealing of one more source of weakness in the traditional type of examination to justify the effort put into them

15

Recommendations—Scaling

THE MARKS of all examiners of a single subject should be scaled to a common standard

All examination marks in different subjects should also be scaled to the same standard

Scaling is the single most important change which should be introduced into the present examination system *now* And it can immediately raise the reliability (in the broadest sense) of examination marks (Misra 1969, as well as this chapter), without any other change whatsoever

It is also the single easiest change It requires no change in the traditional essay type questions It requires no re training of examiners It requires no fresh or special instructions to students It requires no re-education of the public *It requires only an administrative decision*

Scaling was recommended by Mahalanobis as early as 1934 It has been repeatedly recommended, by a variety of writers and commissions, since then Surely forty years is a long enough "educational lag" so that the idea should no longer be considered so radical that it should be postponed, yet again "for further study"

And scaling is so simple that some form of it is used even by elementary school teachers in countries like the U S A Even American school children understand and accept the idea of "grading on the curve", i e the application of the normal distribution curve to the problem of marking examinations The senior author has found both teachers and students in India quite willing to understand and accept the idea, after a simple explanation (see Harper, 1963 a)

We have repeatedly mentioned the need for scaling of examination marks, throughout this book We will not bore the reader by quoting, again, all the various types of evidence which point to this need This chapter will confine itself only to the highlights, and to a brief discussion of some of the practical methods involved

1 What is scaling ?

Your friend says, "The temperature is 40° today" You think, 'It should be very cold—but it is quite hot!' Then you remember that you are old fashioned and prefer to think of degrees Fahrenheit, but your friend is talking about degrees Centigrade So you decide to convert 40° Centigrade into the equivalent number of degrees Fahrenheit

How? You apply the formula $\frac{9C}{5}+32=F$, where F is the degrees Fahrenheit, and C the degrees Centigrade Here it is $\frac{9\times 40}{5}+32=\frac{360}{5}+32$ $=72+32=104°$ No wonder it seems hot—it is 104° today!

Perhaps you used to think that "a degree is a degree" But obviously, a degree Fahrenheit is different from a degree Centigrade The meaning of the word "degree" is different for the two scales *Similarly, the meaning of the word "mark" is different for geometry and for history*—and it is usually even different for the "marks' Professor Ram Das awards in history, and the "marks" Professor Mohan Lal awards in history

'Scaling' is any process which makes the *meaning of the word* "mark" the *same* for Prof Ram Das and for Prof Mohan Lal, and the *same* for both history and geometry

When different examiners of the same subject use different scales, then a student's Division may depend more on the "chance" factor of who his examiner is, then on *his actual merit In the extreme, one student* may get a First while another student of equal merit is declared a Failure just because their answer books happen to be assigned to examiners who use widely different scales As we have seen repeatedly, *each examiner* has, so to speak, his own "personal equation" Scaling simply means finding what this "personal equation" of each examiner is, and correcting the marks which he has awarded, accordingly

The effect of using different scales in different subjects is somewhat different from the above A "mark" of "70 out of 100' in mathematics just does not mean the same thing as it does in, say, English (*Government of India, 1966, p 201*) In English, this mark means, "This is an outstanding student" In mathematics it means, "This student is above average, but not really very brilliant" (In some countries, in fact, a mark of 70 in mathematics would be considered failing)

The scaling methods applied to the marks of different examiners of the same subject are similar to those which are applied to the marks in different subjects But let us discuss the former situation first How can we ensure that the marks of all examiners, marking the same subject, will have the same meaning in terms of the actual merit of the candidates ?

2 Specified distributions of marks

The simplest method of scaling possible is the concept of "grading on the curve" This assumes that (unless there is strong evidence to the contrary) each batch of answer books marked represents an average group Since abilities are assumed to be normally distributed, the marks should be normally distributed A certain proportion of them is, therefore, expected to fall in each of the Classes or Divisions If approximately this proportion does not appear in each Division, then the marks are re adjusted accordingly

To some extent, this is what every examiner does "in his head" The real problem is this No authority has stated or fixed the proportion that should fall in each Division, for a normal group The result of this is that each examiner has his own idea, and hence in practice examiners use a wide variety of marking scales As we have seen (Table 8 3), the average marks awarded the same batch may range from 8 to 22 with equally wide differences in standard deviation

Specifying the distribution of marks expected for an average group—and holding examiners to it—would tend to eliminate at least some of the widest differences between the different scales used by different examiners *And this can be done without any change at all in the present examination system, except for the giving of more adequate instructions*

The instructions of the examining authority to its examiners might be something like this

"We expect that the average mark of all the candidates in this examination will be approximately marks We expect that the pass percentage will be approximately , with a variation of not more than 3% or 4% in either direction We expect that approximately the following percentage will fall in each Division Distinction %, First %, Second % Third %, Fail %

"After marking your answer books, re arrange them in order of merit Then count the number in each Division, and calculate the percentage in each Division and the Pass Percentage If you feel that the batch you have marked is an average or typical batch, then your figures should be very close to the above If they are not, please *readjust*

all total marks proportionately, so that they fit the above *distribution* (You do not need to change the marks on individual questions, only the totals*)

"Some examiners, of course, may have received an extra good batch or an extra weak batch You still should tabulate your results and compare them with the above-mentioned average distribution Then ask yourself seriously, 'Is my batch really so much weaker than the average group as to justify the amount of difference in the marks I have awarded them, from the average distribution specified?' Or, 'Is my batch really as much better than average as the marks I have given them seem to imply?' In most cases, the examiner will find he has to readjust his marks upward or downward If his batch is extra weak or extra strong the distribution does not need to fit the one specified above But it should not be too different '

Initially, instructions may have to be somewhat more detailed than this It is suggested that they be tried out on representative examiners and then clarified as needed A tabulation form can also be provided, to facilitate understanding and application of the general principles

The percentages in each category can be obtained by tabulating a random sample of previous years' results

It may be objected that the entire group of examinees may be stronger or weaker one year than another This may be true, but it is highly doubtful that the differences are usually as much as the variations in pass percentages seem to suggest It is more likely that the variations in pass percentages reflect variations in difficulty of questions, and variations in the scales used by the examiners, at least as much as they reflect any actual variations in achievement (Government of India 1966, p 291) In any case, the flexibility allowed individual examiners in the above instructions will still allow for a year's entire batch being higher or lower than that of a previous year

3 'Raw' marks and "scaled marks

The above will be facilitated if examiners can be taught to think in terms of a two-stage marking process. This is not necessary, but it will help The first marks they assign should always be considered raw, or unscaled marks This, then, frees them to assign these marks in any

* There is a very human tendency, when faced with a major new idea, to throw it out because of some minor defect A predictable number of persons will object to this procedure saying But it makes scrutiny impossible " It does not Scrutiny can still be based on adding up the marks awarded each question to see if the original raw mark is correct Then we can check if the scaled mark assigned this raw mark is the same as for other identical raw marks

way that they like. For example, some may find it simplest to mark every question "out of 5", and then to double the marks of those where the maximum is 10. Much psychological research suggests that the most reliable marking might be to assign a maximum of 7 marks to each question, since seven seems to be the largest number of categories in which human judgement can reliably differentiate anything complex. These raw marks, however obtained, are then added up to produce a total raw mark.

These total raw marks are then translated into scaled marks. Scaled marks are the marks to be reported to the students, and it is these which should fit the distribution of marks specified. If the examiner feels that his whole batch is extra good or extra weak, then he can change the specified distribution—e.g. decide that the average should be 50 rather than 45, or that there should be 3% rather than 1% of Distinctions. Then he assigns scaled marks to his raw marks, so that they will fit this distribution.

The simplest method of converting raw to scaled marks is this: First decide what the highest scaled mark should be, and what the lowest scaled mark should be. The highest raw mark is then set equal to the highest scaled mark and the lowest raw mark to the lowest scaled mark. The remaining marks are distributed proportionately, in between.

A slightly better method has been proposed by Gulliksen (1950), and adapted by Harper (1963 a) to Indian conditions. The senior author and his colleagues have taught this to many Indian teachers, and found that it is easily and quickly understood and accepted.

4. The role of the Head Examiner

The Head Examiner, or Dy. Head Examiner, re-examines a number of answer books from each examiner, chosen at random.

First, let us consider the purpose of this re-examining. Firstly, it seems to be to ensure that each examiner is actually marking the answer books, and not just assigning numbers at random without regard to ment. But this is a minor purpose—if examiners cannot be basically trusted to do their work conscientiously, then we may as well throw out the whole examination system. (As we have seen, there is ample evidence in the studies reported in this book, that *Indian examiners are at least as careful and conscientious in their work as those of other countries.*) A second, and more important, purpose in the Head Examiner's re-examining seems to be to ensure a constant standard among examiners. But we have frequently heard experts complain that the system really does not do what it is supposed to do, i.e. equalize standards. (In the terms we are using in this chapter, we may say that the system does not in practice ensure that all examiners will use the same marking scale.)

Most research studies done in India (see Chapter V), as well as the two reported in this book, amply justify this complaint

Can the situation be improved?

The Head Examiner (or Dy Head Examiner) re examines batches of answer books chosen at random from the whole batch of the work of each examiner Why at random from the whole batch ? A better procedure would be as follows

The Head Examiner selects a few answer books marked as First Division by each of the examiners He then compares these with each other to see if the standard is constant He does similar studies for Distinction, Second Division, Third Division, and Fail In each case, he should select answer books with marks within, say, 5 marks of the middle of the Division

Psychologically, concentrating on a single level of ability at a time is a much simpler task than trying to assess a large number of ability levels, mixed up at random Thus it can be done with a much higher degree of reliability In fact, the Head Examiner will be able to make fairly precise statements as to exactly how much a particular examiner's mark of 50 should be lowered or raised to make it conform to the general standard

If the Head Examiner can wait to begin his work until all of the examiners have finished their work, then he can re-adjust the marks of each examiner himself according to his findings (Or he can specify the adjustments needed, and have a clerk or assistant do it) If—as is more common—the Head Examiner must examine answer books before examiners have finished, and send them back for re examining, then we would suggest this The Head Examiner specifies the number of answer books in each Division to be sent to him He may, for example, say, "Each examiner is to send me 3 answer books with marks between 65 and 70, five with marks between 50 and 55, five with marks between 35 and 40, and five with marks between 20 and 25" He then compares those in each category with each other, and tells each examiner how much (on the average) his marks in each category should be raised or lowered

The above methods will of course have to be adapted to the specific traditions and administrative techniques and problems of each separate examining authority It should be evident, however, that the changes required in present procedures are minimal Yet even these minimal changes can greatly reduce (though not eliminate) the wide variation in standards of different examiners, variations which now make chance—i e the chance of who the examiner happens to be—actually a greater determiner of a student's marks than is his true merit

5 *Randomization and equating*

More sophisticated, mathematically based methods of equalizing scales and standards will, of course, be even more effective than the above method

But to apply more mathematically precise methods, we must meet one of two conditions

(a) Either we must be able to assume that the distributions of marks in the batch of answer books sent to each examiner is approximately the same as the distribution sent to every other examiner ,

(b) Or we must be able to define fairly precisely just exactly what differences exist between the batches sent to different examiners

Randomization It is to meet condition (a) that Dr H J Taylor introduced at Gauhati (Taylor, 1963 c) the system of "randomization of scripts" All answer books for a region are collected in a central office The batches are then thoroughly mixed up From these mixed up answer books, a certain number of answer books is drawn at random and sent to each examiner It is therefore assumed that the distribution of abilities in each batch is approximately equal, i e that the mean and standard deviation of the scaled marks of each batch should be identical

This would seem to be the ideal procedure, for it solves several other problems as well It is virtually impossible for any candidate to know where his answer book will eventually go And no centre's answer books go to one examiner only

However, there are admitted administrative complexities in this system, and for a very large examination they may be overwhelming

If condition (a) cannot be met, then condition (b) can There are several possible ways in which we can define fairly precisely just exactly what differences exist between the batches sent to different examiners

Head Examiner re examines One way grows out of the discussion above, of the role of the Head Examiner The results of his re-examining can be reduced to statistical constants, and applied mathematically to correcting the raw marks of each examiner This would, then, lead to a three stage rather than a two-stage marking process (i) examiner s raw marks, (ii) examiner s scaled marks which he reports to the Head Examiner, and (iii) Head Examiner s final scaling of examiners' marks (Possibly stage ii could be eliminated)

Ordinarily, the Head Examiner re-examines in order to issue instructions to the examiners What we are suggesting here is that he not issue instructions, but rather that he determine the "personal equation' of each examiner and make the necessary corrections himself The actual corrections, of course, can be done by formula, applied by a statistical clerk (The "standard score" method discussed later, would be the easiest to use)

All examiners examine a common batch Harper (1967) has made the following proposal

"A possible solution would be to have the Head Examiner select ten (or, preferably, 20) answer books, representing a wide range of merit (This could probably be done by taking five answer books at random from each of four Centres, centres which were selected to represent a wide variety of expected levels This would save the time otherwise needed to examine the papers before reproducing them The adequacy of the distribution of quality could be estimated quickly by reading only one question in each book) Enough copies would then be made so that each examiner would have a copy of each of these sample answer books (Either the actual hand-writing, etc, could be reproduced by Rotaprinting Xerograph masters, as we did—or the contents could be typed, and copies cyclostyled) Each examiner would then mark these papers along with the rest of his batch The mean and distribution of the marks awarded these sample answer books by an examiner could then be used to determine his "personal equation" They could be compared with the average mean and distribution from all examiners, and the results of all the answer books marked by this examiner scaled to the common standard"

Use of a "link test" A link test is a brief test whose purpose is to assess the level and distribution of ability in the group, rather than to assess any individual within the group Thus it need not have a reliability above 50 or so It is called a "link" test because its only purpose is to link together the scales used by different examiners The link test may be a short objective test of knowledge and understanding of the subject, or it may even be a brief verbal intelligence or (preferably) academic aptitude test In any case, it need not take more than 10-15 minutes of the examination time (Twenty to 30 carefully chosen objective questions are enough) All link tests are sent for scoring by the Board's office, and not to individual examiners The results are not added to any individual student's marks, but are used to determine the range of marks to be awarded to each batch of answer books

After the link test is marked, the raw marks are translated into scaled marks, so as to give the specified proportion in each Division, etc The handiest method is that of Gulliksen (Harper, 1963 a) Or a straight mean and standard deviation transformation can be made, as described later (If the total number of candidates is large, then a random sample of 500 link test results would be quite adequate for this process of defining the scale for the link test) Then the link test results of each examiner's batch are tabulated separately Suppose, for example, the scaled link test results of the batch of candidates assigned to Examiner A, have a mean of

50 and a standard deviation of 10 Then, whatever the raw marks are that Examiner A hands in, they are re scaled so as to have a mean of 50 and a standard deviation of 10 The link test results of Examiner B's batch may have had a mean of 47 and a standard deviation of 12 Then Examiner B s marks are re-scaled, so as to have a mean of 47, and a standard deviation of 12

This may sound complex, but actually it is quite simple, and could be done quite quickly

6 Scaling methods

All scaling methods require first that we specify the scale that we intend to use We may define this in terms of any two or more points and, using the normal distribution curve, build our scale around those points We may define it in terms of the average score and the pass percentage of an average group We may define it in terms of the mean and standard deviation of the group We may define it in terms of the percentage expected to fall in each Class or Division We may even define it in terms of the highest and the lowest score to be awarded any particular group

After defining or describing our scale, we must then decide on its applicability to a particular group If all the candidates are marked by the same examiner, there is no problem—the scale is applied directly to the examiner s results If there is a large number of examiners, but the answer scripts have been randomized (as described in the last section), again there is no problem The scale can be applied directly to the raw marks of each examiner But if the system is such that different examiners can receive batches of different ability, then we need to apply one of the equating methods described above

The next process is the actual conversion of raw to scaled marks There are two general types of methods The first type is based on the mean and standard deviation Many readers will be familiar with "standard scores", though perhaps they have not realized that marks can be converted to *any* arbitrary scale by these methods The second type is called ' equi percentile' or ' normalized" scaling Again, many readers familiar with T-scores or stanines, may not have realized that marks can be converted to *any* arbitrary scale by similar methods*

* It is interesting to note that both T scores and stanines were originally defined as normalized scales McCall originally proposed the T-scale as a normalized scale anchored to the performance of 12 year-olds (hence the "T") The 12-year-old standard has long since been dropped and now there is a growing tendency to use the term "T-scale " for any scale (normalized or standard deviation) with a mean of 50 and a SD of 10 The term ' stanine" is also now used by many, for a standard deviation scale (as well as the original normalized scale) with a mean of 5 and a SD of 2.

7. Standard deviation scaling

This is probably the most widely used scaling method. It has the advantage that it is quick and objective. It is particularly suited to the situations involving "equating" through common batches, link tests, or the re-examining of small samples.

The simple relationship on which standard deviation scaling is based is this:

$$\frac{x-m}{\sigma} = \frac{X-M}{\Sigma} \tag{15.1}$$

where: x is a raw score (i.e. the mark given by any particular examiner), and X is the corresponding scaled mark
m and σ are the mean and standard deviation of the original examiner's marks (i.e. of the raw scores)
M and Σ are the arbitrary mean and standard deviation of the agreed-upon common scale to which the raw marks are to be converted

Applying simple algebraic transformations, we can reduce the above to the form:

$$X = \frac{\Sigma}{\sigma}x + \left(M - \frac{\Sigma}{\sigma}m\right) = ax + b \tag{15.2}$$

A simple example will show the effect of applying this formula. We have earlier used the results of Examiners 15 and 6 (of the ninety marking ten experiment) to show the wide differences between unscaled marks of the same answer books. Let us arbitrarily decide to use Examiner 6's scale as the standard, and convert Examiner 15's marks to Examiner 6's scale. (We could, of course, convert both of them to some other arbitrary scale just as easily.) In this situation, "x, m, and σ" refer to Examiner 15's marks, and "X, M, and Σ" refer to Examiner 6's marks.

It will be remembered that these examiners' marks have the following characteristics:

$$m = 8.1 \qquad M = 22.2$$
$$\sigma = 4\,66 \qquad \Sigma = 6\,52$$

Inserting these statistics in formula 15.2, we obtain,

$$X = \frac{6.52}{4\,66}x + \left(22.2 - \frac{6\,52}{4\,66}8\,1\right) = 1.4\,x + 10.9 \tag{15.3}$$

We can now apply formula 15.3 to each mark awarded by Examiner 15, to find its equivalent mark in the scale used by Examiner 6.

In the following table, the first three columns show the original unscaled marks and the differences between them. The last three columns

show the scaled marks (note that Examiner 6's are unchanged), and the differences between them.

	Unscaled marks of examiner			Scaled marks of examiner		
Candidate	15	6	Difference	15	6	Difference
A	7	23	16	21	23	2
B	5	15	10	18	15	3
C	8	27	19	22	27	5
D	20	34	14	39	34	5
E	11	29	18	26	29	3
F	6	26	20	19	26	7
G	2	16	14	14	16	2
H	7	17	10	21	17	4
I	10	22	12	25	22	3
J	5	13	8	18	13	5
Average difference			14 1			3.9

Let us note, again, that both of these were experienced examiners, and that they were marking exactly the same batch of ten answer books. The average difference between their unscaled marks was 14.1. Scaling reduced this to only 3.9. The *smallest* difference between unscaled marks was 8—the *largest* difference between scaled marks is only 7.

It is obvious that this simple algebraic process can greatly increase the reliance that we can place on any set of marks.

In practice, M and Σ can be defined arbitrarily (based on an analysis of past examination marks). Or they may be defined by the marks of a link test, or of a Head Examiner's re-examining, separately for each batch. Whatever the basis for the scale, the mechanics of converting raw to scaled marks are obviously quite simple. In fact, it is not necessary to apply the formula to each individual answer book. It is simple to set up a table, listing raw marks in the first column and their equivalent scaled marks in the second. (With a calculating machine, such a table can be prepared in ten minutes or less.) Then the examiner, or a clerk, looks up each candidate's raw mark in the left-hand column, reads off his scaled mark, and records it.

8. *Why not just add or subtract a constant?*

One common but crude method of "scaling" sometimes used in India is just to adjust the mean of the marks. If an examiner's mean is above the group mean, the difference is subtracted from each of his marks; if below, the difference is added. While better than nothing, this is quite inadequate. Such an adjustment would be justified if all examiners used the same distribution of marks. However, it is clear from our data that

the differences in distribution (technically called the standard deviation) are nearly as large as the differences in mean Furthermore, *they are more important* When all candidates take the same combination of examina tions (and to some extent even when they take different combinations) *it is the standard deviations, not the means which determine the relative weight of each paper in the total* Therefore, scaling methods which take account of *both* the mean and the standard deviation are needed

The importance of scaling different examiners to the same scale is repeatedly evident in our data Take, for example this pair of mathematics examiners marking the same 50 answer books Their marks correlated highly, $r = 92$

	Mean	Standard Deviation
Examiner D	11 08	8 45
Examiner C	11 08	9 50

Partly because of this difference in standard deviation, the five candidates who were placed in Division II by Examiner D, were spread from Divi sion I to Fail by Examiner C Adjusting the means would have made no difference at all—since the means were identical for the two

Even more striking is this example from history

	Mean	Standard Deviation
Examiner D	18 7	1 96
Examiner C	18 9	5 02

Again the two means are almost identical but the difference in the stan dard deviations can only be described as huge

It is obvious that just a scaling method which only added or sub tracted a constant could have made no difference in these examples Not taking into account the distribution (or standard deviation) as well as the mean or average of each examiner, is like asking a one-legged runner to compete in the Olympics

9 "*Normalized scaling methods*

A big disadvantage of the standard deviation scaling methods is this The scaled marks follow the same distribution as the raw or un scaled marks This would be acceptable if our examiners were trained in objective marking Unfortunately examiners are seldom trained so there are often rather erratic differences between the scales used by different examiners—not only in mean and standard deviation but in the shape of the distribution also It can be shown that the 'steps' in

many examiners' scales are seldom all equal. For example, in our ninety marking ten experiment, we found "gaps" of several marks occurring at different places for different examiners of the same ten answer books. The difference between the highest failing candidate and the lowest passing candidate was as high as 12 marks for some examiners, while it was only 1 mark for others. A more common example is what Taylor has called the J-effect—the tendency to pile up marks just above the passing point. If we then scaled these by mean and standard deviation methods, we would have the absurdity of a J occurring at, say, 41 marks, or some other "meaningless" point. Finally, there is the tendency for lenient examiners to award a disproportionately large number of high marks, and for the stiff examiner to award a disproportionately large number of low marks. While abilities may not be completely normally distributed in some groups, it is unlikely that they are as skewed* as some examiners make them. In other words, forcing a normal distribution on the marks probably does less violence to the true distribution than allowing examiners to impose their own psychological quirks.

All these problems can be avoided *by having examiners report only rank orders and no marks at all.*

The "normalized" scaling methods force the marks into a normal distribution curve. Thus the scaled marks no longer have the same irregular distribution as the unscaled marks—which they would have with standard deviation scaling methods. The normalizing methods accomplish this by specifying that a certain percentage or proportion of the group is to be awarded a particular mark. In stanine scaling, for instance, the top 4% are awarded a mark of 9, the next 7% are awarded a mark of 8, the next 12% a 7, etc. In the familiar T-scale, the proportion awarded each score is smaller. Percentile ranks are another "normalized" scale with which many teachers are familiar. The top 1% of the group have a percentile rank of 99, the next 1% have a percentile rank of 98, etc.

The same thing can be accomplished just by deciding that, for a group of a given size, a rank (order of merit) of 1 is equivalent to a mark, say, of 80. A rank of 2 may be equivalent to a mark of 78. In the middle of the scale, the ranks will, of course, tend to bunch up. Ranks of 100, 101, 102, 103, and 104 may all be equal to a mark of, say, 50.

The marks equivalents of the various ranks can be decided on the basis of tabulating past examination results Or they may be set arbitrarily, with reference to the normal distribution curve. (A statistician can translate ranks into any score system, taking account of the fact that

* This chapter is addressed both to the non-statistician and to the technical expert Therefore, some technical statements are unavoidable The educational administrator, teacher, or parent should just ignore these technical statements, as they are not essential to his understanding of the general principles involved

there is always a wider difference between ranks 1 and 2 than between, say, ranks 49 and 50, out of 123 ranks.)

The examiners, then, report only ranks or order of merit. They may, of course, mark their answer books in any way that they wish, but in the end they must put them in order of merit and just report this order. They should also be restricted, say, to declaring not more than three answer books as "tied" or of "equal rank". The ranks are then converted into marks in the Board's office.

Rank order scaling is very easy to apply when the answer books have been randomized. (In fact, it is probably easier than any other method, including the adding or subtracting of constants to equalize the means.) Where one of the "equating" methods has been used, the process becomes a little more complex. A set of conversion tables will have to be prepared, for use with batches with different means and standard deviations. But any competent statistician can do this. He only has to remember that rank orders are a non-linear scale (like the more familiar percentile scale) which he is converting to scaled marks, which are a linear scale.

10 The advantages of rank order scaling

Examiners who are asked to place answer books in rank order, or order of merit, are asked to do a task which is far simpler for human judgement than is the assigning of absolute marks. They may, of course, begin by assigning some sort of marks. But they are relieved of the necessity of going back and changing their marks, when they discover the question is easier (or harder) than they thought, and therefore they have been marking too high (or too low). They are also relieved of the agonizing decision about a borderline candidate, "Should I pass him or should I fail him?" They merely put him in his rank position, and let the statistics make the decision. Again, in a situation of internal assessment, it is far easier for a teacher to justify a particular rank assigned a student in his class, than to answer the angry parent who says, "You could have given him a First if you had wanted to". Would not the parent be much less likely to say, "You could have declared my son to be better than Ram Das, if you had wanted to?"

In Chapter III, Section 5, and again in Chapter VIII, Section 5, we examined the difference between marks and rank orders. We found that examiners agree on ranking the best candidate, but they do not agree on how many marks to award him. Thus if the examiners award only rank orders, *but the marks are awarded by the Board's office*, there will be a much higher degree of reliability in the marks of the candidates. The absurd situation in which the same candidate can be awarded a

Distinction by one examiner and a low Fail by another is likely to be eliminated.

To demonstrate this, let us convert the rank orders in the ninety marking ten experiment, to marks. There are several methods by which we could convert such rank orders to marks. We might do it on the basis of some arbitrary marking scale. Or we might find the average mark which each of the 90 examiners assigned his top candidate, the average mark each assigned his Number 2, etc. But perhaps it is best to take our clue from the "correct marks" of the various candidates. (Remember, of course, that this is an unusual situation. We do not know "correct marks" in any public examination. The *specific* type of conversion shown here, therefore, would not be applicable in ordinary practice.)

Let us list the ranks and correct marks of the ten answer books (see Table 8.6).

	Mean Rank	Correct Mark
D	1.2	27 0
E	2.1	23 1
C	3.3	19 0
I	4 9	15.5
F	4 9	15.4
A	5 4	14 5
B	7.7	11.2
H	7.8	10 9
G	8 6	9 8
J	9 2	8 8

We could, of course, just assign Rank 1 a mark of 27, Rank 2 a mark of 23, Rank 3 a mark of 19, etc. But for a more exact solution, we draw a graph. On the ordinate we list the marks (from 7 to 28), and on the abcissa the ranks 1 to 10 We plot our points, and connect them with a curve. From this graph, then, we read off the following rounded values:

Rank	Mark	Rank	Mark	Rank	Mark
1	28	4 5	16	8	11
1.5	26	5	15	8 5	10
2	24	5 5	15	9	9
2 5	22	6	14	9 5	8
3	20	6 5	13	10	8
3.5	19	7	12		
4	17	7.5	11		

Table 3.4 may now be re-written, in terms of scaled marks rather than of rank orders. (We have used the more detailed data in Table 8.5, which included ties as half-ranks.) See Table 15 1.

TABLE 15 1

THE EFFECT OF RANK ORDER *SCALING ON THE MARKS* DISTRIBUTIONS IN TABLE 3 1

Scaled Marks	Answer Books D	E	C	J	F	A	B	H	G	I	Total
28	70	10	1		1						82
26	7	6	1		—						14
24	12	55	11	2	1						81
22	—	4	3	1	—						8
20	—	12	43	11	8	2		1			77
19	1	—	6	2	3	2		—			14
17		—	11	20	18	13		—			62
16		2	7	8	14	12	1	—			44
15		1	5	19	24	25	3	5	1		83
14			1	13	10	19	4	6	3	1	57
13			—	6	3	6	7	7	3	—	32
12			—	1	3	6	20	10	5	4	49
11			1	6	3	4	28	27	23	11	103
10				1	1	1	9	10	9	5	36
9					—		8	16	17	17	58
8					1		10	8	29	52	100
Total	90	90	90	90	90	90	90	90	90	90	900

It is interesting to compare Table 15 1 the scaled marks with Table 3 1 the unscaled marks Perhaps the simplest comparison is just in the range of marks which are awarded each answer book For convenience we are reproducing below the lowest and highest mark and the range for both the unscaled and the scaled marks

	Unscaled	*Scaled*
D	17–35 19	19–28=10
E	11–38=28	15–28=14
C	8–30 23	11–28=18
I	7–26=20	10–24=15
F	6–26=21	8–28=21
A	7–23=17	10–20 11
B	5–17=13	8–16= 9
H	3–19 17	8–20 13
G	2–17=16	8–15= 8
J	2–15 14	8–14= 7

Scaling of marks, by the rank order method, has reduced the range of marks awarded each candidate by an average of one-third. With unscaled marks, some candidates were placed in as many as five different Divisions—now only four are placed in more than two divisions, and none in more than three. The extremely high marks have been eliminated, and so have the extremely low marks. Obviously, many of the idiosyncracies of individual examiners have been greatly moderated or even eliminated by this simple process. One can place much more reliance on scaled than on unscaled marks. We have eliminated a major source of difference between examiners. We have substantially reduced the very large "chance" factor in traditional essay-type examinations.

11. Scaling of marks in different subjects to a common standard

All that we have said concerning the scaling of marks of different examiners of the same subject, also applies to the scaling of the marks of different subjects to a common standard. The need for scaling examination marks has been pointed out repeatedly (Mahalanobis 1934, Harper 1963, Taylor 1963). The data in this study repeatedly pointed to this need.

Perhaps the most striking example of the need for scaling was the difference in the mathematics scale and other scales, as has already been pointed out. This resulted in nearly as wide average differences between geometry examiners (in a supposedly "exact" subject) as between other examiners. It resulted in a much wider standard error of measurement, despite higher inherent reliability, for mathematics than for other subjects.

As we have pointed out earlier, it is the standard deviation, not the mean or even the "maximum marks" which determines the weight a particular subject carries in the aggregate marks. Mathematics examiners, using a scale with a standard deviation twice that of any other subject, are thereby *making mathematics twice as important as any other subject* in determining a student's rank in the aggregate marks.

An example should make this clear. From Table 4.5, we can see that the top 5 geometry candidates were awarded (in the Original marking) 50 marks out of 50. But the top history candidate was only awarded 36 marks, the second, third, and fourth 31 marks, and the fifth only 30 marks. Let us make a table of some of the high ranks:

	Marks awarded in	
	History	Geometry
Rank 1	36	50
„ 5	30	50
„ 10	28	49
„ 15	27	48

Suppose Candidate A ranks First in history and 15th in geometry His total marks are 84 Candidate B is obviously superior, for he ranks First in geometry and only 10th in history Yet his total marks are only 78 Because marks are not scaled, the inferior candidate receives the higher total marks, just because he happens to be strong in geometry A similar example could be worked out at the bottom of the scale to show that here the student who is weak in geometry is penalized more than the student, of equal ability, whose weak subject happens to be history

Adding together marks of such widely differing scales is about as rational as adding inches and centimeters, or degree Fahrenheit and degrees Centigrade

Let us close this chapter by again reminding the reader The scaling of examinations requires no change in the current type of essay examination, it requires no re education of examiners it requires no fresh instructions to students it requires no re-education of the public. Since scaling is to be done in the Board s or the university s office, all that is needed *is an administrative decision* to apply a modern and thoroughly tested method, for the immediate improvement of the reli ability of examination marks Taylor et al (1966 b) put the matter in a nutshell

> 'The notoriously large uncertainties of university examining arise, not so much from the unreliability of the examiners, as from the failure to [scale] When scripts are divided amongst many examiners, their marking standards must be brought to the same level by scaling If this is not done, the examiner who passes 15% of his scripts is put on the same footing as the one who passes 85%, with the effect that poorer candidates pass and better ones fail'

16

Recommendations—Question-wise Multiple Marking

ONE *obvious method to increase the reliability of marks* is to have two or three (or more) examiners evaluate each answer book Since this doubles or triples the work involved, this is very expensive

A second, less obvious, method is to have each question evaluated by a different examiner Since this does not increase the number of examiner-hours involved (each question is read only once), it adds little or no expense to current methods

The ninety marking ten data gave us an ideal basis for studying these two methods

Not only is the second method cheaper—it is also better

These studies strongly support the contention that the reliability of the marks can be increased substantially, if each candidate's work is marked by several different examiners, each examiner marking only one question

1 Examiners who agree are likely to be wrong

First let us discuss a practice which is widely used in higher degree examinations Two examiners are appointed, and they examine every answer book The answer books on which their marks agree are taken to

be correctly marked Only the answer books on which they disagree are submitted to a third examiner

This seems to be, of course, "common sense" Unfortunately, "common sense" is often scientifically incorrect Dr H J Taylor and Mr L N Tluanga have written a booklet entitled *The Problem of the Third Examiner* (1964) They point out, and prove mathematically, that when examiners disagree their average is more likely to be correct than when they agree!

For those who may not be able to follow Taylor and Tluanga's mathematical argument, the following very simplified version may help

Suppose we have two examiners, both of equal ability and experience

Examiner A marks one-third of his papers too high (let us call this +1) He marks one third too low (let us call it —1) The remaining third are marked approximately correctly (let us call this 0, for "zero error') Examiner B marks with the same degree of error

Examiners A and B mark the answer books *independently* This means that each type of mark by Examiner A has an equal chance of being paired with each type of mark of Examiner B Thus, + 1 for Examiner A is paired with + 1, 0, and —1 for Examiner B Zero (0) for A is paired with + 1, 0, and —1 for B A's —1 is paired, successively, with B's + 1, 0, and —1 Thus there are nine possible ways in which Examiners A and B can mark a given set of answer books

Now let us arrange these 9 ways in order so as to bring out the salient points

Examiner		
A	B	Average Error
+1	+1	+1 0
0	0	0
—1	—1	—1 0
+1	—1	0
—1	+1	0
+1	0	+ .5
0	—1	— .5
0	+1	+ .5
—1	0	— .5

The two examiners *agree* in only the first 3 out of 9 cases Of these three averages, *two are quite wrong*, one being too high (+ 1 0) and the other being too low (—1 0)

The two examiners *disagree widely* in 2 out of the 9 cases—and in both these cases, *the average is exactly right*

In four cases, the two examiners *disagree slightly*. In all four, the average is only *slightly wrong*.

When the examiners disagree slightly, the average is closer to the "correct mark" than the average is in two out of the three complete agreements.

SUMMARY: When two examiners award *different* marks, the average is more likely to be correct, or nearly correct, than it is when they award the same mark.

CONCLUSION: There is *no* justification for the practice of assigning a third examiner to only those answer books where the two marks given a candidate *differ* by 5, 10, or 15 points. If anything, the third examiner should be appointed when the two previous examiners *agree*.

RECOMMENDATION: We have been talking about two experienced examiners. And we have been talking in terms of simple "agreement" and "disagreement", without taking into account *how much* they disagree.

It is certainly true that two good examiners will disagree with each other by fewer marks, on the average, than will two poor examiners. Two poor examiners may have several disagreements of 10 marks or more—two good examiners may have no disagreements of 10 marks or more.

It is therefore recommended that: If the two examiners disagree by 10% marks or more *in more than one-tenth of the answer books*, then *the entire batch of answer books* (those on which they agree as well as those on which they disagree) *should be sent to a third examiner*. The third examiner should mark the answer books independently of the first two, and the final mark should be *the average of the three examiners*.

Note that Vernon (1940) has recommended that there should always be three examiners, and the data given in this paper support this view. In India, however, this may not be economically feasible. The method proposed allows the advantages of three examiners where necessary, while requiring only two when a third is not really needed.

2. How many examiners are needed ?

By drawing at random from our ninety marking ten data, we can test the effect of having the marks of one, two, three, etc. up to ten examiners averaged. We have the advantage that we already know the correct mark of each answer book. This is the average mark awarded by 90 experienced examiners. Therefore, we can compare the average marks of two, three, etc. examiners with the correct mark. This tells us how much error remains in the average.

Table 16.1 presents this study. Answer book A had a correct mark

(the average of the 90 examiners) of 14 5 At random, the first examiner drawn for this answer book was Examiner 80 He awarded this answer book 10 marks His error, therefore, was 10—14 5=—4 5 marks The second examiner drawn at random was Examiner 32 He also awarded 10 marks The average of these two examiners, 10, still was —4 5 marks in error Examiner 70 awarded 17 marks The average of the three examiners then became 12 3, and the error of this average was reduced to —2 2

TABLE 16 1

AVERAGE MARKS OF FROM TWO TO TEN EXAMINERS COMPARED WITH THE CORRECT MARK (MEAN OF 90 EXPERIENCED EXAMINERS) FOR EACH OF TEN HISTORY ANSWER BOOKS

Candidate		Examiner 1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Correct Mark
A	Examiner	(80)	(32)	(70)	(26)	(71)	(98)	(19)	(18)	(43)	(42)	
	Mark	10	10	17	9	19	11	15	14	18	18	
	Average		10 0	12 3	11 5	13 0	12 7	13 0	13 1	13 7	14 1	14 5
	Error	—4 5	—4 5	—2 2	—3 0	—1 5	—1 8	—1 5	—1 4	—0 8	—0 4	
B	Examiner	(90)	(81)	(34)	(92)	(68)	(95)	(44)	(11)	(54)	(96)	
	Mark	9	13	14	12	13	8	11	9	9	10	
	Average		11 0	12 0	12 0	12.2	11 5	11 4	11 1	10 9	10 8	11 2
	Error	—2 2	—2	8	8	1 0	3	.2	—1	—3	—4	
C	Examiner	(20)	(16)	(80)	(93)	(71)	(55)	(16)	(57)	(73)	(12)	
	Mark	22	22	21	24	20	26	22	21	19	12	
	Average		22 0	21 7	22.3	21 8	22.5	22.4	22.2	21 9	20 9	19 0
	Error	3 0	3 0	2 7	3 3	2 8	3 5	3 4	3.2	2 9	1 9	
D	Examiner	(66)	(63)	(42)	(7)	(2)	(19)	(81)	(36)	(94)	(19)	
	Mark	30	22	25	32	25	34	33	23	25	34	
	Average		26 0	25 7	27 3	26 8	28 0	28 7	28 0	27 7	28 3	27 0
	Error	3 0	— 1 0	—1.3	3	—2	1 0	1 7	1 0	7	1 3	
E	Examiner	(43)	(47)	(36)	(64)	(10)	(25)	(14)	(55)	(48)	(2)	
	Mark	25	16	26	38	21	28	26	23	27	20	
	Average		20 5	22 3	26 3	25 2	25 7	25 7	25 4	25 6	25 0	23 1
	Error	1 9	—2 6	— 8	3 2	2 1	2.6	2 6	2 3	2.5	1 9	
F	Examiner	(14)	(55)	(97)	(48)	(22)	(88)	(13)	(66)	(34)	(59)	
	Mark	19	12	21	19	15	20	17	19	18	10	
	Average		15 5	17 3	17 7	17.2	17 7	17 6	17 8	17 8	17 0	15 4
	Error	3 6	0 1	1 9	2 3	1 8	2 3	2 2	2 4	2 4	1 6	
G	Examiner	(87)	(56)	(66)	(76)	(21)	(70)	(2)	(75)	(88)	(41)	
	Mark	7	8	12	4	12	15	10	10	12	4	
	Average		7 5	9 0	7 7	8 6	9 7	9 7	9 7	10 0	9 4	9 8
	Error	—2 8	—2 3	— 8	—2 1	—1 2	— 1	— 1	— 1	2	—4	
H	Examiner	(29)	(52)	(84)	(89)	(35)	(8)	(72)	(30)	(56)	(46)	
	Mark	10	12	12	12	8	15	9	9	5	9	
	Average		11 0	11 3	11 5	10 8	11 5	11 1	10 9	10 2	10 1	10 9
	Error	—.9	1	4	6	— 1	6	2	nil	— 7	— 8	
I	Examiner	(20)	(88)	(98)	(72)	(45)	(37)	(58)	(54)	(36)	(57)	
	Mark	25	17	12	17	11	11	12	12	15	17	
	Average		21 0	18 0	17 7	16 4	15 5	15 0	14 6	14 7	14 9	15 5
	Error	9 5	5.5	2 5	2.2	9	nil	— 5	—.9	— 8	— 6	
J	Examiner	(43)	(8)	(22)	(70)	(10)	(42)	(10)	(55)	(43)	(64)	
	Mark	9	12	11	12	8	10	8	13	9	7	
	Average		10 5	10 7	11 0	10 4	10 3	10 0	10 4	10 2	9 9	8 8
	Error	2	1 7	1 9	2.2	1 6	1.5	1.2	1 6	1 4	1 1	

If you look over Table 16 1 you will find several examples of Taylor and Tluanga's thesis Where examiners agree (like Examiners 80 and 32 for Answer Book A) their average is *frequently quite wrong* But look, for example, at the first two examiners of Answer Book F. They disagree widely (19 and 12), but their average is almost exactly correct In general, when examiners disagree their average is more likely to be correct than when they agree Remember that the examiner reliability for these data was 83

Table 16 2 summarizes the results of our multiple marking It shows the number of times errors of each size occurred when the results of two, three, four, up to ten examiners were averaged The bottom row shows the average error The average error for a single examiner was 3 2 When we used two examiners, we cut this error by about one-third Three examiners cut it by one-half But *it required averaging the results of ten examiners, to cut the error from 3 2 to 1 0* It required at least five to reduce the error to 1 3, on the average

TABLE 16 2

SUMMARY OF TABLE 16 1 SHOWING THE NUMBER OF TIMES A DIFFERENCE OF EACH SIZE OCCURRED, ON EACH RE EXAMINING, BETWEEN THE AVERAGE OF TWO TO TEN EXAMINERS AND THE CORRECT MARK (NOTE DIFFERENCES TABULATED WITHOUT REGARD TO SIGN MARKS ARE OUT OF 50)

Size of difference from correct mark	Examiner 1	2	3	4	5	6	7	8	9	10	Total *frequency*
9 9 9	1										1
8 8 9											
7-7 9											
6-6 9											
5 5 9		1									1
4-4 9	*1*	*1*									2
3 3 9	3	1		3		1	1	1			10
2 2 9	2	2	3	4	2	2	2	2	3		22
1-1 9	1	2	3		5	3	3	3	1	5	26
0-0 9	2	3	4	3	3	4	4	4	6	5	38
Average Error	3.2	2 1	1 5	2 0	1 3	1 4	1 4	1 3	1 3	1 0	

3 Each examiner marks a different question

It is doubtful that any Board will appoint ten examiners for each answer book—or even five If it takes, say, a half an hour to mark an answer book, then five examiners would take two and a half hours—increasing the cost of marking by five

But suppose, instead, each examiner reads a different question?

Then we have five examiners, yet the total time taken in examining an answer book remains only half an hour

Harper (1963 c) has proposed that the candidates answer each question in a separate answer book Each examiner, then marks only one question (The article cited gives many practical suggestions as to how this might be carried out as well as theoretical discussions of its value)*

The ninety marking ten experiment provides data to test Harper s proposal At random we drew Examiner 14 We assigned him Question 1 Thus the marks which Examiner 14 had assigned to Question 1 in each answer book were written down as that answer book s marks for Question 1 Then at random we drew Examiner 24 and assigned him the second question The marks he had assigned that question in each answer book, were recorded as that answer book's marks for that question Similarly, examiners 36, 41, and 48 were assigned the remaining three questions Thus, Answer Book A's total mark is made up of the mark Examiner 14 assigned to it on the first question, plus the mark Examiner 24 assigned to it on the second question etc

We repeated this experiment four different times, each time drawing 5 different examiners and assigning a single question to each of them

Table 16 3 shows the results of these four experiments For example, Answer Book A s correct mark is 14 5 The total of the first five examiners' marks was 15 5, an error of + 1 0 The total of the second five examiner's marks was 10 5, an error of —4 0 The total of the third five examiners marks was 12 5, an error of —2 0 The last set of five examiners awarded a total mark of 16 5, an error of 2 0

It is interesting to compare these results with those in Table 16 2 First we may tabulate the errors of different sizes for ease in comparison

4-4 9	3
3 3 9	2
2 2 9	7
1-1 9	15
0 0 9	13
	—
	40

* The Vice-Chancellor of Kanpur University Mr Radha Krishna proposed at a meeting at which the senior author was present a similar but in some ways simpler scheme All answers would be written in one book as at present Several examiners would have to sit together The first examiner would mark only the first question answered in every book He would then leave the answer book open so that the next examiner could mark the next question Since each examiner is still marking several different questions (i e all those questions which happen to come first or second etc) this is not so theoretically sound as Harper s proposal On the other hand it has the advantage that it could be applied immediately without any change in present answer book procedures

About one-third of these are errors of less than one point, which is comparable to the results achieved with four or five examiners reading each entire answer book. In general, the above distribution is slightly better than when two examiners read each entire answer book—i.e., when twice as much work is done as in Harper's method.

The bottom line of Table 16.3 gives the average error for each of the four experiments. (We might, of course, have done many more such experiments—but the average errors were so close, that it did not seem necessary to replicate further.) These average errors range from 1.37 to 1.69. All of them are lower than the average error of a pair of examiners in Table 16.2; and half of them are lower than the average for three or four examiners.

It is evident that it takes at least three examiners reading each entire answer book, to get results which are more reliable than those obtained by having each question read only once but by a separate examiner. Since the latter method costs only half as much (in time and examiner's fees) as having two examiners evaluate each entire answer book, it is *definitely the preferred method. Even if* we can afford to have three examiners evaluate each entire answer book, it is still questionable if the slight gain in accuracy would be worth the extra cost.

TABLE 16.3

MARKS OF CANDIDATES COMPARED WITH CORRECT MARKS WHEN A DIFFERENT EXAMINER MARKS EACH OF THE FIVE QUESTIONS, EACH EXAMINER MARKING THE SAME QUESTION IN ALL 10 ANSWER BOOKS. (NOTE: ALL CANDIDATES ANSWERED THE SAME 5 QUESTIONS).

	Four sets of 5 examiners chosen at random								Correct Mark
	First		Second		Third		Fourth		
Candidate	Mark	Error	Mark	Error	Mark	Error	Mark	Error	
A	15.5	1.0	10.5	—4.0	12.5	—2.0	16.5	2.0	14.5
B	10.5	— .7	10	—1.2	10	—1.2	10	—1.2	11.2
C	20	1.0	19	0	16	—3.0	20	1.0	19.0
D	24.5	—2.5	28.5	1.5	26.5	— .5	28	1.0	27.0
E	23.5	.4	23.5	.4	22.5	— .6	23	— .1	23.1
F	16.5	1.1	12.5	—2.9	14.5	— .9	16	.6	15.4
G	10.5	.7	8.5	—1.3	5.5	—4.3	8	—1.8	9.8
H	6.5	—4.4	9	—1.9	10.5	— .4	7.5	—3.4	10.9
I	16.5	1.0	13	—2.5	15.5	0	13	—2.5	15.5
J	7	—1.8	10	1.2	8	— 8	6.5	—2.3	8.8
Mean		1.46		1.69		1.37		1.59	

Thus it is both cheaper and better to have each question read by a separate examiner The results of each candidate are then based on as many examiners as there are questions. Apparently at least some of the examiners individual idiosyncrasies can be averaged out by this simple method

17

Recommendations—General

The last two chapters have dealt with specific recommendations which would improve marks without any change whatsoever in the present pattern of examinations. Each of them would require only an administrative decision, to introduce scaling, and to introduce question-wise marking of answer books.

But with examination reform "in the air", it is hoped that we will go much further.

Twenty-two years ago the *University Commission Report* (1950)—commonly known as the Radhakrishnan Commission—pointed out that the traditional examination is "invalid . . . inadequate . . . subjective and therefore not reliable", and recommended (p. 329) "the introduction of . . . objective examinations in the universities of India at the earliest possible time". Perhaps the first step is for someone to decide what "the earliest possible time" is.

But short of such radical change, there is much that can be done to improve the more familiar type of examination. One is tempted to discuss the many proposals made: More freedom to paper-setters to determine the most valid method of assessment (instead of being required always to follow past patterns)—better questions—better training of examiners—use of multiple forms of an examination to frustrate copying —more experimentation with "open book" examinations—use of machines (and even carbon-backed forms!) to free examiners from the unnecessary paper work which reduces reliability, because it reduces

the time available for examining—etc , etc , etc But we will confine ourselves to six recommendations which arise directly out of these two studies

1. Six recommendations

1 *Eliminate options* The data show clearly that if no choice of questions is allowed, then both examiner and content *reliability*—and, hence, marks reliability—are increased significantly Options must be allowed when there are alternative textbooks, as in languages, of course Otherwise, as all experts agree, they are indefensible

2 The elimination of options will be more acceptable if *the number of questions is greatly increased* The traditional, long essay type question undoubtedly elicits—i e requires for its answering—such skills as clear thinking, organization, creativity, style, etc But whether it actually *measures* these skills to any reliable degree is highly doubtful There is probably nothing valuable measured by a long answer that cannot be measured just as well by a one or two-page answer It is possible to set as many as 20 one page questions in a three-hour examination The present "short notes' type of question provides a precedent, though the form needs to be greatly improved, dropping the present emphasis on factual memory, and substituting more questions that say "compare", "contrast", "what is the difference between", ' discuss the relationship of to ", "outline the essential features of ', etc Many of the present "short notes ' questions could actually be answered quite adequately in a sentence or two—which suggests a liberal use of even shorter answers "My experience with Air University demonstrated that good questions requiring answers of one or two sentences in length showed as high reliability and higher item validity coefficients than items in objective tests given under the same time limit, even though the objective tests contained about 50%—75% more questions " (Warren G Findley, personal communication)

3 Even with two page, one page, or one sentence answers, *efforts must be made to improve marking instructions, and training of examiners*, so that inter-examiner reliability can be raised to the 90 which Gulliksen (1950) recommends as *minimal* This means nearly doubling in real terms (i e in terms of Fisher's z) most of the examiner reliabilities found in this study Better marking instructions is the minimum first step But training of examiners is also needed If any examiner feels he needs no training, let him do the experiment of marking the same 100 answer books twice If the correlation coefficient is less than 90 he has lost his case

4 We have seen that many examinees are falsely labelled ' Failed",

when actually it is the examination system which has failed *Many of these injustices can be eliminated by using the grace marks system proposed by Taylor and Tluanga* (1963 *a*) Unlike the traditional rather arbitrary methods, Taylor's system is based on the reality of examiner (and test) error, and therefore makes a rational (rather than purely arbitrary) and fully fair correction (See Appendix B for a version of Taylor's paper which may be somewhat more comprehensible to the non-mathematician than the original) Incidentally, Taylor and Tluanga's tables are based on the standard error of marking But tables based on the much wider standard error of measurement could easily be derived

5 *Routine statistics should be calculated after every examination* At the very least, a Board should know the mean, standard deviation, percentage in each Division, and content reliability of every examination which it conducts Probably some sort of item analysis should also be carried out And if possible, a few answer books should be re-examined to determine examiner reliability, marks reliability, and measurement error The relative uniformity of the results in our studies suggest that these statistics can be derived quite adequately from a rather small sample of scripts in each subject, say 500 at the most This is particularly true if a stratified random selection system is used—first selecting a specified number of examiners at random, and then selecting a specified number of answer books at random, from the work of each

6 Finally, if examination reform is ever to become a reality, *widespread publicity and education of both teachers and the general public is needed Talented journalists should be induced to study the many findings of research, and to reduce them to simple articles for the popular press* It is a mistake to assume that the scientists and educators who are competent to do the research are necessarily the best ones to explain it to the public Despite extensive research, there is still an appalling lack of knowledge and understanding even among teachers and educational administrators, as to the extent of examination unreliability, the reasons for it, and the well developed technology already available to correct it The task is an educational one, and in dealing with the layman it is often the competent journalist who is really the best educator

2 In conclusion

The question raised by the Secretary of the Board of Secondary Education, who provided the answer books and examiners' addresses for these studies, was this "Are the variations within permissible limits" ?

What are the "permissible limits" ?

The psychometrician would say that the "permissible limits" are the greatest degree of precision which can be obtained, within the goals,

aims, purposes, and financial resources of any particular examination

Are there any values in the present methods of constructing administering marking, and scaling examinations which are *great enough* to justify the almost complete lack of precision in the end product—the marks assigned the student ?

This is a question that cannot be answered by the psychometrician It is a question which can only be answered by the educational authorities in charge of the system

The Education Commission (Government of India *1966*, pp *290-291*) hit the nail on the thumb, *so to speak*, when they said,

> " what is lacking is not knowledge, but will, courage and perseverence "

APPENDIX A

Content Reliability

Abstract and example from Harper's "Internal consistency Reliability of Traditional type Examinations when a Choice of Questions is Allowed'

Ebel s intra class correlation (see Guilford, *Psychometric Methods*, 2nd Ed , pp 383-385) uses analysis of variance to derive the reliability of a set of ratings This same approach can be used for essay type examinations when no choice of questions is allowed, and is equivalent to a generalized form of Kudar Richardson formula 20 The analysis of variance may be reduced to an algebraic formula as follows

$$r_{kk} = 1 - \frac{kr\Sigma X^2 + (\Sigma X)^2 - r\Sigma(\Sigma X_r)^2 - k\Sigma(\Sigma X_k)^2}{(k-1)\,[r\Sigma(\Sigma X_k)^2 - (\Sigma X)^2]}$$

where, r = number of rows, i e the total number of examinees in the table

k = number of columns, i e the number of questions to be answered by each examinee

X = the mark on any single question

ΣX^2 = the sum of the squares of all the X's (Each X must be squared and all squares added)

ΣX = the sum of all the X's

$(\Sigma X)^2$ = the square of the sum of all X's (Be sure to distinguish this from ΣX^2 The latter requires that each X be squared first, and all these squares summed, the former that all X s first be added and the sum be squared)

$\Sigma(\Sigma X_r)^2$ = the sum of the squares of the row-sums (The sum of each row, i e the total mark of each candidate, is squared, and these squares are added)

$\Sigma(\Sigma X_k)^2$ = the sum of the squares of the column sums. (The total of all the marks awarded each question is squared, and these squares are added)

Harper's paper adapts this to several types of cases where a choice of questions is allowed in a traditional-type examination. In the simplest case, the candidate has a free choice of questions (e.g. "answer any 5 out of the 9 questions"), and all questions carry equal marks. It is shown that in this case the examiner assumes that all questions are interchangeable, and therefore are of equal difficulty; and therefore the formula reduces to the following form:

$$r_{kk} = 1 - \frac{kr\,\Sigma X^2 - r\,\Sigma(\Sigma X_r)^2}{(k-1)\,[r\Sigma(\Sigma X_r)^2 - (\Sigma X)^2]}$$

The following example is designed both to illustrate the method, and to show the effects of choice on essay-type examination reliability. The difference in item difficulty is *deliberately exaggerated* to demonstrate these effects.

Let us assume there are six candidates, A, B, C, .. F. The examination has 7 questions, out of which each candidate is to answer 4 only. The questions differ in difficulty. Let us further assume that we know how each candidate might have scored if he had answered *all* of the questions.

The first table shows the difficulty of each question for each candidate, and also his marks if he chose the four best questions.

	Questions							Marks if chose best
	1	2	3	4	5	6	7	
Average Difficulty	8	7	6	5	4	3	3	
Marks of Candidate A	10	9	8	7	6	5	5	34
B	9	8	7	6	5	4	4	30
C	8	7	6	5	4	3	3	26
D	7	6	5	4	3	2	2	22
E	6	5	4	3	2	1	1	18
F	5	4	3	2	1	0	0	14

Notice that, on each question in this artificial example, the candidates are ranked in order of ability. The reliability of this examination, in this form, is obviously perfect. The examiner has made no assumptions about the difficulties of the questions, when he requires all candidates to answer all questions.

But when he allows a choice, the examiner apparently does so on the assumption that the questions are interchangeable, that is they are of equal average difficulty—and that each candidate will choose the questions on which he thinks he can do best. But we know that examiners cannot

judge difficulty accurately—and that candidates cannot always guess how examiners will mark a particular question So, when a choice is allowed, the situation is likely to be like the next table And this table is used to illustrate the calculation of reliability Each candidate is allowed to answer "any 4 out of the 7 questions" Notice that the ranks of the candidates are changed *because of their choices* (Gayen's studies as well as Harper's have shown that this does happen)

	Questions							ΣX		
Candidates	1	2	3	4	5	6	7	Marks	$(\Sigma X_i)^2$	Square each mark and sum the squares
A				7	6	5	5	23	529	
B		8	7	6	5			26	676	The sum of the column sums = the sum of the marks This provides a check, both here and with the total of the frequency distribution below
C			6	5	4		3	18	324	
D	7	6	5			2		20	400	
E		5	4	3	2			14	196	
F	5	4	3	2				14	196	
	12	23	25	23	17	7	8	115	2321	
								ΣX	$\Sigma(\Sigma X_i)^2$	

Prepare a frequency distribution of the marks awarded to each question

X	f	fX	fX^2
8	1	8	64
7	3	21	147
6	4	24	144
5	7	35	175
4	3	12	48
3	3	9	27
2	3	6	12
	24	115	617
	kr	ΣX	ΣX^2

$$r_{kk} = 1 - \frac{(4)(6)(617) - 6(2321)}{(4-1)[(6)(2321) - (115)^2]}$$

$$= 1 - \frac{14\,808 - 13\,926}{3(13\,926 - 13\,225)}$$

$$= 1 - \frac{882}{2103}$$

$$= 1 - 419$$

$$r_{kk} = 58$$

APPENDIX B

A Scientific Substitute for Grace Marks

The present grace marks rules of the Board are admittedly rather arbitrary, and I have been asked by the Secretary of the Board whether they can be placed on a more scientific basis.

Yes, I believe they can.

What follows is based on a paper by Dr. H. J. Taylor entitled "Grace Marks" and published by the U.G.C. in a booklet called *Three Studies in Examination Technique*. (Those who wish a fuller explanation than is given below, may refer to the original paper.) Dr. Taylor is an eminent British scientist (physicist) and educator, who has spent most of his life in service in India. He published this paper while he was Vice-Chancellor of Gauhati University.

No measurement is perfectly accurate. There is a "margin of error" in every kind of measurement. If we ask ten men to measure the length of a table to the nearest millimeter, we will probably receive ten different answers. Even so important a "physical constant" as the speed of light is known only within an accuracy (called "probable error") of plus or minus 100 meters per second.

The purpose of an examination is to *measure* the degree of knowledge, understanding, etc. possessed by the student. Like *all* measurements, an examination mark is subject to error. This is no reflection on the integrity of the examiners. It is the fundamental nature of measurement. As every scientist knows, *measurement without error is impossible.*

Actually, there are two major sources of error in examination marks: (a) if the candidate had been asked different questions, his marks might have been different; and (b) if his answer book had been marked by a different examiner, his marks might have been different.

The mark which the candidate would have obtained if there were *no* measurement error in examination marks may be called his "correct mark". (This mark is, of course, unknown—and, unfortunately, unknowable.)

It seems reasonable to state that:

1. If a candidate's "correct mark" is probably below 33% in a subject, then he should fail in that subject.

2 If a candidate's "correct mark" is probably 33% or above, he should pass in that subject—regardless of what mark has actually been awarded to him by the examiner

3 A candidate should not be failed if his marks are below 33% *due to measurement error*

If these three statements are accepted, it remains only to give a numerical value to the word "probably' and the problem is solved

1 Relationship between "correct mark" and mark awarded by examiner

Unfortunately, we cannot know the "correct mark"—any more than we can know the "true speed" of light, or the "true length" of a table.

But we do know something about the relationship between the "correct marks" of a group of candidates, and the marks actually awarded by their examiners

Assuming that the examiner is awarding each candidate the mark which he honestly feels is correct, then the inevitable "errors of measurement" will be distributed according to the laws of chance Once we have determined the average measurement error, then the laws of chance will tell us the exact *probability* that the mark awarded will differ from the "correct mark" by any given amount

2 Determining the size of "measurement or marking error"

The physicist is lucky He can determine "measurement error' by measuring the speed of light, or the length of the table, hundreds of times But we cannot administer hundreds of examinations to each candidate! So we must find a different method

If the same set of answer books is marked by two examiners, and they award different marks to many of them, then these differences are obviously due to "measurement error" By applying standard statistical techniques to these differences, we can determine the distribution of 'measurement error" for that examination

Ideally, the "standard error of marking' (as it is called, technically) should be determined separately for each paper, and for each examiner When the Board has an electronic computer, perhaps this can be done For the present, however, we have several Indian research studies to draw on those of H J Taylor, A K Gayan, A E Harper, Jr It is fair to say that a minimum estimate of the average "standard error of marking" is 5% of marks (It is at least this large in almost all subjects—and in some subjects it is much larger)

Applying standard statistical theory we can state that 2/3 of all candidates will receive a mark which is within 5% of marks above or below their "correct marks" Ninety-nine per cent of all the "measurement errors" will be within about 13% of the marks awarded

What does this mean? It means that, in a single subject, 1/3 of the candidates whose "correct mark" is 33 will be awarded marks which are either 28 marks or lower, or are 38 marks or higher (See Supplement A for details) To put it differently, 1/6 of all the candidates with 28 marks or lower *should have passed,* because their "correct mark" is 33% Of course, it is also true that 5/6 of these candidates have "correct marks" of 32% or lower, and therefore are correctly failed

Unfortunately, we cannot know *which* candidate has been failed by "measurement error" and which has been properly failed This is because "measurement errors" are randomly distributed, i e according to the laws of chance

3 "Measurement error" in an examination

Fortunately, when the candidate answers several papers, these errors tend somewhat to cancel each other out Thus the chance of a candidate being failed "by error" in two subjects is far less than the chance of such failure in one subject As every mathematician knows, in such cases we multiply the probabilities (Non-mathematicians may see Supplement B for a demonstration) Thus if the probability of a mark of 28 in one subject is 1/6, the probability of marks of 28 in two subjects is only 1/36 ($1/6 \times 1/6 = 1/36$)—and the probability for marks of 28 in all subjects, for a candidate whose correct mark is actually 33% in each of them, is only 1/46,656

Obviously, we should have no hesitation in failing a candidate if there is only this microscopic chance—one in forty-six thousand—that he should have passed However, if there is one chance in six that we are doing an injustice in failing him, then he seems to be a candidate for "grace marks"

4 Setting the standard

But between these extremes, where shall we draw the line? At what level of "probability that he should have passed" do we pass a candidate? At what level of probability that his "correct mark" is 33% do we say, "This probability is so small that we must ignore it, and fail him "?

Dr Taylor has proposed an eminently logical solution If a candidate has exactly 33% marks in each of his six subjects, we pass him without question Calculate the probability that such a candidate's "correct mark" is actually 33% or above in every subject This probability is then the minimum level of probability required for passing the examination

If the candidate obtains a mark of 33% in a particular subject, then we find from statistical tables that the chances are 53 98 out of 100 that his "correct mark" is 33% or better (The chances are a little better than 50% because a "correct mark" of 32 5 would be considered a 33) Thus the

probability for a candidate who obtains exactly 33% in each of 6 subjects is 53.98/100 multiplied by itself six times, or $(.5398)^6$. According to Dr Taylor's calculations, this comes to 2 47 out of 100

If the probability that a candidate's "correct mark" is 33% or above in each of his six subjects is 2 47/100 or higher, he should be considered as having passed the examination. (The level would be different for 5 or 7 subjects.)

5. *What is likely to be the effect of using this system ?*

It is worth quoting Dr. Taylor on the usual situation :

"If a candidate's true level is just above the pass line in all subjects, and his mark is underestimated in only one subject, he fails the examination But the converse is not true. If a candidate's true level is just below the pass line in all subjects, his mark must be overestimated in all of them before he can pass. The second contingency is much less likely than the first The number of good students who fail is therefore larger than the number of bad students who pass ... There can be little doubt that this is a contributory cause of the abnormally high failure rates in the examinations of many Indian universities"

In other words, the abnormally high failure rate is probably due to "errors of measurement", and if we can reduce the effect of these inevitable errors, we will increase (and *legitimately* increase) the pass percentage.

6. Summary

The usual grace marks system condones failure. The system proposed by Dr. Taylor *does not condone ANY failure on the part of the candidate.* What it does is to *correct for measurement error*, and ensure that the candidate probably will not fail "just by chance".

We pay no attention to the *number* of subjects failed. The number is irrelevant, if the "failures" are due to "measurement error". However, the mathematics of the system shows us that the probability of failing several subjects "by chance" is *very* much less than the probability of failing only one or two "by chance", i e. through "error of measurement".

We pay no attention to the aggregate marks as such. *No candidate with an aggregate of less than 33% will pass.* However, the number of marks "in excess" in the aggregate is *not* directly relevant to the probabilities of "measurement error", as such.

The existence of these "errors of measurement" is no reflection on the examiners concerned They are inherent in the very process of any measurement. It is, therefore, proper that we should take them into account in deciding whether a candidate "passes" or "fails".

The likelihood of a good candidate being failed by "measurement error" is *much greater* than the likelihood of a poor candidate being passed by "measurement error". Therefore this system, which takes scientific account of "measurement errors" and allows for them, is likely to raise the pass percentage of the Board. And it is proper that it should do so, because the high failure rate was partly due to errors of measurement.

7. How to calculate "passing probability"

First we will describe the detailed method for those who "want to know what they are doing". However, this detailed method is *not* recommended—it is much easier to use Dr. Taylor's table. So those who are not interested in the details can go right on to the next section.

(1) Subtract 33% from each of the candidate's marks. Suppose, for example, the candidate's marks are 36, 33, 30, 34, 32, 36. We do it this way:

Marks obtained:	36	33	30	34	32	36
Subtract	33	33	33	33	33	33
	+3	0	−3	+1	−1	+3

(2) Find the "probability" for each of these differences. If you want to calculate these yourself, you can calculate each $(X-32.5)/\sigma$ (remember that $\sigma=5$), and look up the p value in a table of the normal curve. Then multiply the six p values by each other, to obtain the total probability.

Deviation	+3	0	−3	+1	−1	+3
Probability	.76	× .54	× .31	× .62	× .46	× .76

If you multiply these out, you will obtain a total probability of .027576 or 2.76/100. Since this is better than our standard of 2.47, the candidate has Passed.

NOTE: Under usual "grace marks" rules, this candidate would have failed, since he "failed" in two subjects and earned no grace marks. However, it has been shown that his failure is *due to measurement error*, not due to his own lack of ability. He is actually a better candidate (as is shown by his higher "passing probability") than the candidate who receives exactly 33% marks in each subject.

8 Simplified calculations using Dr. Taylor's Table

Dr H. J. Taylor has provided the attached Table which eliminates the need for multiplication. The multiplications have been done in the Table. It therefore requires only a few seconds to find the "passing probability" of any candidate.

(1) Subtract 33% from each of the candidate's marks. Suppose, for example, the candidate's marks are 36, 33, 30, 34, 32, 36 We do it this way:

Marks obtained	36	33	30	34	32	36
Substract	33	33	33	33	33	33
Deviations	+ 3	0	− 3	+ 1	− 1	+ 3

(2) Find the first deviation in the top of the Table and read off the number immediately below this In the line below +3 you will find 76 (This is the probability of a single deviation of +3)

Next find this Probability in the left hand column of the Table Then go across this row until you reach the column of the second deviation In our example We find the row numbered 76 Going across it to the 0 column we find the number 41 (This number is the product of the two probabilities 76× 54= 41)

Next find this new Probability in the left hand column of the table. Then go across this row until you reach the column of the third deviation In our example We find the row number 41 Going across to column—3 we find the number 13 (This is the product of our three probabilities so far 76× 54× 31= 13)

We continue this process until we have multipl ed by all the deviations Dr Taylor suggests that we avoid the last few rows of the table where numbers are small and therefore rounding errors are larger It is recommended that as soon as an entry less than 10 is reached it should be multiplied by 10 to carry the calculation back to the top of the table *All* the steps in our example are as follows

Row 100 column +3=76
Row 76 column 0−41
Row 41 column −3−13
Row 13 column +1= 8
(Multiply 8×10−80)
Row 80 column −1−37
Row 37 column +3−28

We divide our final answer by 10 (to correct for previously multiplying by 10) and find our answer to be 2 8/100

(3) If this final value is greater than the minimum passing probability of 2 47 the candidate passes If the final value is less than 2.47 the candidate fails (This cand date passes His failures in two subjects were due to errors of measurement)

9 Notes

1 With a little practice these calculations can be done in a few seconds using this Table A sl ding scale can be designed to help find the proper deviation in any given row Subtraction of 33 from each mark

can be done in the head, without writing it down.

2. The descriptions above are based on an examination with 6 subjects with 100 marks maximum for each. Where the number of papers or the number of marks are different, the necessary tables and method can easily be worked out.

SUPPORTING MATERIALS IN SUPPLEMENTS

A. Distributions of error.
B Why probabilities are multiplied.

Supplement A—Distributions of error

Suppose the "correct mark" of a candidate is actually 33%. Suppose his examination answer book is marked by, say, 100 experienced examiners. What sort of marks will he receive?

Some examiners will award exactly 33—but most will not.

Many will award marks of 34, or 35, or 36.

Many will award marks of 32, 31, or 30.

Fewer will award marks above 38; but this same number will award marks below 28.

The marks will be distributed "at random" above and below the "correct mark".

Actually, if we take our "standard error of marking" as 5 marks, and our "correct mark" as 32.5 (which is the lower limit for 33%), the distribution will be as follows. On the average:

1	examiner	will	award	46	marks
2	examiners	will	award	43-45	marks
15	,,	,,	,,	38-42	,,
28	,,	,,	,,	34-37	,,
8	,,	,,	,,	33	,,
28	,,	,,	,,	29-32	,,
15	,,	,,	,,	24-28	,,
2	,,	,,	,,	21-23	,,
1	examiner	will	award	20	marks
100	examiners				

Notice that 46 of these 100 examiners have awarded "failing" marks to this "passing" candidate. But 54% of the examiners have actually passed him. Thus the chances are slightly better (54:46) that a candidate whose "correct mark" is 33% will be awarded a passing than a failing mark. However, we cannot easily afford to ignore the fact that 46% of the examiners failed a candidate who should have passed.

Notice that we could re-write this table on the basis of a "correct

mark" of 32 (or, more exactly, 31 5)—i e a candidate who *should* fail It will look like this

1	awards	45	marks
2	award	42-44	marks
15	„	37-41	„
28	„	33-36	„
8	„	32	„
28	„	28-31	„
15	„	23-27	„
2	„	20-22	„
1	awards	19	marks

This candidate, whose "correct mark" should fail him, is failed by 54 examiners—but is passed erroneously by 46 examiners.

Dr Taylor's "passing probability" method takes into account *both* the probability that a "pass" mark is awarded to a "fail" candidate, as well as the probability that a "fail" mark is awarded—due to chance *errors of measurement—to* a "pass" *candidate*

If these results look extreme, consider the following table. This is based on *actual data* Copies of the same Higher Secondary School *History examination answer book were sent to 90 experienced examiners* The marks out of 50 awarded by these examiners were as follows

2	examiners	awarded	23-26	marks
9	„	„	20-22	„
28	„	„	17-19	„
22	„	„	14-16	„
18	„	„	11-13	„
9	„	„	8-10	„
2	„	„	6-7	„

The average mark awarded by these 90 experienced examiners was 15 4 The candidate's "correct mark ' is clearly a failure—yet 39 of *these experienced examiners* passed him, two of them in the Second Division *If you will consult chapter III where* these data are published you will find that some answer books received an even wider range of marks

Supplement B—Why probabilities are multiplied

Why do we multiply the probabilities by each other (rather than using some simpler process like adding)? The following demonstration will show how this works

Suppose we have a coin. It has two sides, named Heads and Tails Let "H" mean Heads, and "T" mean Tails

If we toss the coin, there are two ways in which it can land on the table:

(1) H
(2) T

If we call the total "1", then it is obvious that the chance of the coin coming up Heads is 1/2 and the chance of its coming up Tails is 1/2.

Now suppose we toss two coins. They might arrange themselves in any of the following ways:

(1) H H
(2) H T
(3) T H
(4) T T

We saw that the chance for one coin coming up Heads was 1/2. Count the chances for two coins to *both* come up Heads. We find just one of the four patterns shows two Heads—so the chance of two Heads is 1/4 Notice that 1/4 is $1/2 \times 1/2$.

Let us toss three coins. There are now eight ways in which they might arrange themselves. Our theory of multiplying chances says that the probability of *three* heads is $1/2 \times 1/2 \times 1/2 = 1/8$ (i e. 1/2 multiplied by itself three times). Is this correct?

(1) H H H
(2) H H T
(3) H T H
(4) H T T
(5) T H H
(6) T H T
(7) T T H
(8) T T T

APPENDIX C

The Problem of the Crossed-out Tick Marks

It is in the nature of scientific research that it is an exploration into the unknown And it is in the nature of the unknown that the problems to be encountered in its exploration cannot always be clearly foreseen

One such problem which was not adequately foreseen was the possible effect of the first examiner on the second It was not foreseen because the planners did not realize that so many lines, ticks, and crosses had been placed in the answer books When this was discovered, about two months of extra labour went into trying to disguise them

Was the disguising effective?

We have seen that, in comparison with studies in other countries, our examiners seem to have somewhat higher examiner reliability than the average We have discussed the several reasons why Indian examiners might be expected to have higher examiner reliability than, say American examiners. However, the nagging doubt remains Was the disguising of previous marks effective? Or did the previous marks spuriously raise examiner reliability?

Four lines of evidence have been gathered to explore this question

1 A comparison with other studies done in India should have been useful Unfortunately, few complete parallels were found The subjects involved differed, the experimental technique differed, etc. The most nearly parallel is the recently completed study of the junior author (Misra 1970), in which his means are based on two completely indepen dent markings, as against our ten of questionable independence His examiner reliability for history had a mean of 71, as against our median of 74 for 'Other re-examining His mathematics mean was 81 as against ours of 95 Thus if there is any influence of the previous marks, it is not in the paper which is considered to be most subjectively marked Taking these facts along with comparisons with other studies, a conservative statement is this Our examiner reliabilities *may* be slightly higher than

for somewhat similar studies done in India But no more positive statement than this can be made

2 We asked the examiners A questionnaire was sent out on several aspects of the experiment, and 36 of the 40 examiners sent in their replies Half of them claimed to recognize the answer books that they had previously marked—but when asked how many of the 100 these were most of their answers were so far away from 50 (the correct number) as to suggest that they did not really recognize their own previous work When asked, "Do you think that your re marking was influenced by the previous marks in the answer books?" none marked "Yes, very much ", 3 marked ' Somewhat influenced ", and 33 marked "Little or no influence "

3 At our suggestion, the examiners used green ink in marking the second time We inspected quite a few answer books, checking visually for the relationship between the green and red marks (Remember there were red crosses added where there had been ticks, ticks added where there had been crosses, and extra ticks, crosses, and underlinings added at random) Most of the green x's and ticks were, of course, near the old red x's and ticks But some green marks were found where there were no red marks Most important, there were quite a few places where there were red marks but no green marks It still could be that the existence of a red mark (even though disguised, so the examiner could not tell if the original mark was favourable or unfavourable) called the examiner's attention to the possible existence of a point to be marked However, he certainly used his own judgement as to whether or not to mark it

4 In the most laborious process of all, we selected a sample of four sets of 50 answer books each for each subject They represented the "Same' re-examiner with the highest examiner reliability, the "Same" with the lowest examiner reliability, the "Different" re examiner with the highest examiner reliability, and the "Different" with the lowest examiner reliability On the fifteen to twenty five pages of each of these 800 answer books, we counted every red X and underline mark (The number of ticks and of X's was identical, except for the earlier ones where we had tried to disguise by crossing out the tick marks) We hoped that this would give us correlation figures that we could use, with partial correlation techniques, to correct reader reliabilities for the influence of the previous examiners

Unfortunately, the results of this study were ambiguous Had we foreseen the problem, and counted the marks before disguising and then re-counted them after, we could have obtained more clearly interpretable answers

5 A similar study was done for the ninety marking ten data

Our overall conclusion is just this There may have been some effect from the fact that the original markings, even though disguised,

could not be removed from the examination books Ideally, of course, completely independent marking would require that the first examiner leave no marks in the answer book This factor may have raised examiner reliabilities by as much as, say, five correlation points (05) But we cannot say positively that it did, and anyway the difference is negligible Even if the examiner reliabilities were raised by the fact that the examiners were not "completely independent', they are still so low that no serious reliance can be put on the resulting marks for differentiating accurately among individual candidates

APPENDIX E

The following information, not included in this book, is available in computer print-out form. If any scholar wishes to do further calculations, a copy may be obtained on loan from: Director, Bureau of Educational Research, Ewing Christian College, Allahabad, U.P., 211003, India.

Sums, sums of squares, and sums of cross-products for the three variables correlated: marks of first examining, marks of second examining, and number of pages in answer book, separately for each batch of 50

Standard errors of the means, and standard errors of the differences for uncorrelated marks, for each batch of 50.

Standard errors of the standard deviations, and standard errors of the differences, and their significance for uncorrelated standard deviations, for each batch of 50.

Standard deviations of the differences in marks between pairs of examiners, for each batch of 50.

Also hand calculations for the two formulas for the significance of the difference between correlated standard deviations.

Anyone reading between the lines of this book will realize that certain other calculations are also available. Since we do not have duplicates, we do not list them here. If you desire anything special, for research purposes, however, please write.

Bibliography

AMERICAN PSYCHOLOGICAL ASSOCIATION *Publication Manual of the American Psychological Association* 1967 *Revision* Washington D C American Psychological Association, 1967

AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION & NATIONAL COUNCIL OF MEASUREMENT IN EDUCATION *Standards of Educational and Psychological Tests and Manuals* Washington D C American Psychological Association, 1966

ANASTASI, A *Psychological Testing* New York Macmillan Company, 1961

ANASTASI, A Improving Construction and Use of Educational Tests In A Anastasi (Ed) *Testing Problems in Perspective* Washington, D C American Council on Education 1966

BALLARD P B *The New Examiner* London University of London Press, 1923

BASUMALLIK T Reliability of Essay Type Examination A Review of Studies *Journal of Education and Psychology*, 1959, *17*, 127 156

BHANOT, I V *Wastage and Stagnation in the Faculty of Medicine and Technology of the M S University, Baroda* Baroda Faculty of Science, M S University, Baroda, 1961

BORA K P Some Observations on Grace Marks *Journal of the University of Gauhati* Science 1963 *16*

BOSE, P K & CHOUDHURY, S B Scaling Procedure in Scholastic and Vocational Test *Sankhya* 1955, *15*, 197

BRACHT, G H & HOPKINS K D Objective and Essay Tests Do they Measure different Abilities? Paper presented at a joint session of the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, 1968

BUROS, O. K. *Summary of Definitions, Basic Concepts, and Formulas in the Schematization of Test Reliability.* Highland Park, N.J.: The Gryphon Press, 1963, (Hectographed).

CHATTERJEE, K. R. Some Observations on Examination Reform in Indian Universities. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

CHATURVEDI, M. S., & CHATURVEDI, V. Examination Reform in Higher Education in India. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

CHAVAN, B. H. Examination Reform. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

CHITRAKARA, R. S. Wastage and Retardation in Education. *Indian Journal of Educational Administration and Research*, Autumn, 1961, 24-32.

CIESLAK, E. C., COWLES, J. T., & DRAGOSITZ, A. M. *Examinations in Indian Higher Education.* New Delhi: U.S. Information Service, 1959.

COFFMAN, W. On the Validity of Essay Tests of Achievement. *Journal of Educational Measurement*, 1966, *3*, 151-156.

COREY, STEPHEN M. Improving Teacher Education: Testing and Evaluation. *Examination Abstracts*, 1966, *3*, 34-35; abstracted from *Naya Shikshak*, 1964, 1-5.

CRONBACH, L. J. *Essentials of Psychological Testing.* New York: Harper and Brothers, 1960.

CURETON, E. E. Validity. In E. F. Lindquist (Ed.), *Educational Measurement.* Washington, D.C.: American Council on Education, 1951.

CURETON, E. E. The Average Spearman Rank Correlation When Ties are Present: A Correction. *Psychometrika*, 1965, *20*, 377.

D'ANGEL. Observations on Reform of Examinations. *Examination Abstracts*, 1964, *3*, 31-32, abstracted from *Progress of Education*, 1964, *38*, 310-312.

DANDEKAR, V. M. Statistical Reform of the System of Evaluating Results of Mass-Conducted Examinations. *Indian Educational Review*, 1968, *3 (1)*, 1-19.

DATTA, SUDHIR. *Application and Standardisation of the Holzinger-Crowder Unifactor Tests.* Unpublished Doctor of Philosophy (Science) thesis submitted to the University of Calcutta, 1967.

Dave, R. H., & Patel, P. M. *Analysis of Results of Board Examinations.* New Delhi : Department of Curriculum and Evaluation, National Council of Educational Research and Training, 1966.

DAVE, R H *Review of Research in Educational Evaluation* New Delhi 16 Department of Curriculum and Evaluation, National Council of Educational Research and Training 1968

DAVIS F B Item Selection Technique In E F Lindquist (Ed) *Educational Measurement* Washington, D C. American Council on Education 1951

DAVIS, F B *Item Analysis Data—Their Computation, Interpretation, and Use in Test Construction* Cambridge, Massachusetts Graduate School of Education, Harvard University, 1956

DEPSE. *Analysis of Results of the Secondary School Examinations in India (March 1961)* New Delhi 16 Directorate of Extension Programmes for Secondary Education, National Council of Educational Research and Training 1962 (a)

DEPSE *Progress of Examination Reform in India* New Delhi 16 Directorate of Extension Programmes National Council of Educational Research and Training 1962 (b)

DEPSE *Optional Questions in Tests and Examinations* 1963

DEPSE. *Private Candidates and Secondary School Examinations A Study* New Delhi-16 Directorate of Extension Programmes for Secondary Education, National Council of Educational Research and Training New Delhi, undated (a)

DEPSE *A Study of Failures in Secondary Schools* New Delhi 16 Directorate of Extension Programmes for Secondary Education, National Council of Educational Research and Training New Delhi undated (b)

DESAI, D M Examination Reform in Indian Universities An Action Programme Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai 1970

DESHMUKH, A G *Wastage and Stagnation in College Education* Nasik Road Government of India Press 1963

DESHMUKH, A G, & KAMAT, A R *Facts about Fergussonians* Undated

DIEDERICH, P B Exercise Writing in the Field of Humanities, 1957 In A Anastasi (Ed) *Testing Problems in Perspective* Washington, D C American Council on Education, 1966

DUBOIS, P A Test Dominated Society China *1115 B C.—1905 A D*, 1964 In A Anastasi (Ed) *Testing Problems in Perspective* Washington, D C American Council on Education, 1966

EBEL, R L The Frequency of Errors in the Classification of Individuals on the Basis of Fallible Test Scores *Educational and Psychological Measurement*, 1947, 725 734

EBEL, R L, & Darmin D E Tests and Examinations 1958 In C. W Harris (Ed) *Encyclopedia of Educational Research* New York Macmillan Company, 1960

EBEL, R L Must All Tests Be Valid? *American Psychologist, 16,* 1961, 640 647

EBEL, R L *Measuring Educational Achievement* New Delhi Prentice-Hall of India, 1965

EBEL, R L How An Examination Service Helps College Teachers to Give Better Tests, 1953 In A Anastasi (Ed), *Testing Problems in Perspective* Washington, D C American Council on Education, 1966

EDGEWORTH, F Y The Statistics of Examinations *Journal of Royal Statistical Society*, 1888, 599 675

ENGELHART, M D Exercise Writing in the Social Sciences, 1957 In A Anastasi (Ed), *Testing Problems in Perspective* Washington, D C American Council on Education, 1966

ELEHANCE, D N Examination Reform at the University Level Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon Madurai, 1970

EXAMINATION COMMITTEE *Report on Examination Reform* New Delhi. University Grants Commission, 1962

FRENCH J W *Schools of Thought in Judging Excellence of English Themes* Princeton, New Jersey Educational Testing Service, 1961

FLETCHER F M Internal Assessment *Journal of Education and Psychology, 1961, 18,* 441-451

FREDERIKSEN N In Basket Tests and Factors in Administrative Performance 1960 In A Anastasi (Ed), *Testing Problems in Perspective* Washington, D C American Council on Education, 1966

FOSTVEDT, D R Criteria for Evaluation of High School English Composition *Journal of Educational Research* 1965, *59,* 108 112

GARRET, H E *Statistics in Psychology and Education* Bombay Vakils, Feffer and Simons, 1965

GAYEN A K , NANDA, P B MATHUR R K , DUARI P , DUBEY, S D , & BHATTACHARYA N *Measurement of Achievement in Mathematics* Kharagpur Indian Institute of Technology, 1961

GAYEN, A K , NANDA, P B DUARI, P , BHATTACHARYA, N , MUKHERJEE, H N & MATHUR R K *Measurement of Achievement in English* Kharagpur Indian Institute of Technology, 1962

GAYEN A K , NANDA, P B DUARI, P , BHATTACHARYA, N , MUKHERJEE, H N , & MATHUR R K *Measurement of Achievements in Some Indian Languages*, Kharagpur Indian Institute of Technology, 1966

GAYEN A K , NANDA P DUARI, P BHATTACHARYA, N , & MUKHERJEE, H N *Measurement of Achievement in Physics and Chemistry* Kharagpur Indian Institute of Technology, 1967

GAYEN, A K Examination and Evaluation in Higher Education Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai 1970

GEORGE, A *Study on Assessment of Standard of Valuation and Comparative Performance of Colleges in the Pre-University Examination* Kerala Department of Statistics, Kerala University, 1964

GODBOLE, N W *Examination Reform* Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970

GOLAY, W H Working Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon, 1970

GOVERNMENT OF INDIA, *Report of the Education Commission* 1964-66, *Education and National Development* New Delhi Ministry of Education, 1966

GUILFORD, J P *Psychometric Methods* Tokyo Kogakusha Company, 1954

GUILFORD, J P *Fundamental Statistics in Psychology and Education* New York McGraw-Hill Book Company, 1956

GULLIKSEN, H *Theory of Mental Tests* New York John Wiley and Sons, 1950

GULLIKSEN, H Methods for Determining Equivalence of Measures *Psychological Bulletin*, 1968, *70*, 534 544

HARPER A E, Jr Why Neglect the S E M ? *Psychological Studies*, 1959, *4*, 1-13

HARPER, A E Jr Recent Advances in Psychometry *Silver Jubilee Volume of Vidya Bhawan Society*, Masulipatam 1960, 1-15

HARPER, A E, Jr Chance in the Traditional Examination *Journal of Education and Psychology*, 1962, *20*, 100-118

HARPER, A E, Jr, GUPTA, B, DAS, & SANGAL, S P *Item Analysis Chart and Instructions* (2nd ed) Delhi Mansayan 1962

HARPER, A E, Jr Marking Objective Examinations *Teacher Education*, 1963, *8*, 10-17 (a)

HARPER, A E, Jr Objective and Traditional Examinations Some Research *Education and Psychology Review*, Spring 1963, 63 78 (b)

HARPER, A E Jr Essay Type Examinations A Proposal *The Journal of University Education*, 1963, *2*, 33-117 (c)

HARPER, A E, Jr The Use of Objective Examinations in India Paper presented to the first conference of the Danforth Indian Fellows, 1963 (Cyclostyled) (d)

HARPER, A E, Jr Down with the Validity Coefficient *Journal of Vocational and Educational Guidance*, 1965 *3* 76-86

HARPER A E, Jr *Essay Examination Reliability* Unpublished MSS, 1966 (a)

HARPER, A. E., Jr. *Examiner Reliability: A Study of Ninety Experienced Examiners Marking the Same Ten Answer Books.* Unpublished MSS, 1966. (b)

HARPER, A. E., Jr. Ninety Marking Ten—A Study of Examinations. *Indian Educational Review,* 1967, *2,* 26-41. (a).

HARPER, A. E , Jr. Review of Aptitude Tests for Mathematics and Statistics. *Sankhya.* 1967, 219-222. (b)

HARPER, A. E., Jr., MISRA, V. S., & GUPTA, R. L. *Four Thousand Answer Books Re-marked.* Unpublished MSS, 1967.

HARTOG, P., & RHODES, E. C. *An Examination of Examinations.* New York: The Macmillan Company, 1935.

HILL, WALKER, H. How to Improve Essay Examinations. *Teacher Education,* 1963, *8,* 8-9.

HILL, WALKER, H. Examinations for Secondary Education. *Education and Psychology Review.* Baroda, January, 1964, 4-8. (a)

HILL, WALKER, H. Examination Improvement and Examination Reform. *Journal of Education & Psychology,* 1964, *22,* 63-70. (b)

HILL, WALKER, H. What Research Says about Examinations. *Teacher Education,* 1965, *9,* 1-10.

HILL, WALKER, H. Problem of Internal Assessment. *Rajasthan Board Journal of Education,* 1966, *4,* 1-8. (a)

HILL, WALKER, H. Should Students Be Exposed to Errors. *The Progress of Education,* October 1966, 1-4. (b)

HILL, WALKER, H. The Teaching of English in India. *The Progress of Education.* November, 1966, 1-4. (c)

HILL, WALKER, H. How Examinations Influence Methods of Study. *Rajasthan Board Journal of Education,* January 1967, 3-9. (a)

HILL, WALKER, H. Malpractices in Examinations. *The Progress of Education,* 1967, 1-6. (b)

HILL, WALKER, H. The Relation of Evaluation to Wastage in Education. *Rajasthan Board Journal of Education.* April 1967, 3-9. (c)

HILL, WALKER, H. The Progress of Examination Reform (Part I). *The Progress of Education,* April 1967, 1-7. (d)

HILL, WALKER, H. Review on the Measurement of Writing Ability. *The Progress of Education,* April 1967, 1-2. (e)

HILL, WALKER, H. The Progress of Examination Reform (Part II) *The Progress of Education,* June, 1967, 1-8. (f)

HILL, WALKER, H. Why Examination System Needs to be Changed. *Journal of Education & Psychology,* 1967, *25,* 20-27. (g)

HILL, WALKER, H. The Role of the Training College in Examination Reform. *Teacher Education,* 1967, *1,* 21-28. (h)

HOSSAIN, Q. M. Standardisation of Examination Marks. *Sankhya,* 1940-41, *5,* 295-298.

HUBBARD, J P. *Programmed Testing in the Examinations of the National Board of Medical Examiners* Paper presented at the Invitational Conference of Testing Problems of Educational Testing Service, N J 1963

HUDGON, L Degree Class and Attainment in Scientific Research *British Journal of Educational Psychology*, 1960, *51*, 67 73

IYER, T S K V *Examination Reform* Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970

JAIN, P. C *Internal Assessment in Higher Secondary Schools* *Educational India* 1962, *29*, 165-168

KADIYAL, R A Examination Reform Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970

KAMAT, A R, & DESHMUKH, A C Prediction Value of Marks at the Preceding Examination in Forecasting Marks at the Succeeding Examination *Artha Vijnan*, 1961, *3*

KAUL ADALATI, A N Examination Reform Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970

KELLEY, E L Alternative Criteria in Medical Education and Their Correlates Paper presented at the Invitational Conference on Testing Problems of Educational Testing Service, Princeton, N J, 1963

KLEIN, S P, & HART, F M Chance and Systematic Factors Affecting Essay Grades *Journal of Educational Measurement*, 1968, *5*, 197-206

LEHTOVAARA, A The Predictiveness of the Finish Matriculation Examination A Differential Psychological Study *Psychological Abstract*, 1966, *40*, 10454 (Abstracted from *Annales Academiae Scientiarum Fennical*, 1966, B-140)

LELE, P T, PATEL, P M, PARIKH, N P, & PALKAR, S G *Empirical Studies in Examinations* Baroda Examination Reform and Research Unit, M S. University, Baroda, March 1962 (a)

LELE, P T, PATEL, P M, PARIKH, N P, & PALKAR, S G *Qualitative Analysis of Essay Questions Research Studies in Examinations* Baroda Examination Reform and Research Unit, M S University, Baroda, December, 1962 (b)

LELE, P T, PATEL, P M, PARIKH, N P, & PALKAR, S G *Empirical Studies in Examinations* Research Monograph 2 Baroda Examination Reform and Research Unit, M S University, Baroda March 1963

LELE, P T, PARIKH, N P, PALKAR, S G, & PARIKH R C. *Empirical Studies in Examinations* Research Monograph 3 Baroda Examination Reform & Research Unit, M S University, Baroda, March 1964

LELE, P T, PARIKH, N P, PALKAR, S G, & PARIKH, R C *Empirical Studies in Examinations* Research Monograph-5 Baroda Examination Reform & Research Unit M S University Baroda, March 1966

LIDDLE, SUSHIL K V *Development and Standardisation of a Scholastic Aptitude and Developed Ability Test for High School Classes* Unpublished Ph D thesis Agra University, 1965

LINDVALL C M *Testing and Evaluation An Introduction* Harcourt Brace and World, 1961

LINDQUIST, E F *A First Course in Statistics* Boston Houghton Mifflin Company, 1942

LOKNATHAN, S Working Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970

MAHALANOBIS P C & CHAKRAVARTY, K. N Analysis of Marks in the School Leaving Certificate Examination in the United Provinces, India, 1919 *Sankhya*, 1934, *1*, 245 266

MAHAJANI, G S Examination Reform Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon Madurai, 1970

MALLICK, A K A Study of Relationship Between Internal Assessment and Secondary Board Examination Marks *Journal of Education and Psychology*, 1968, *16*

MICHAELS W J, & KARNES, M R *Measuring Educational Achievement* New York McGraw-Hill Book Company, 1950

MINISTRY OF EDUCATION, GOVERNMENT OF INDIA (ED) *Report of the Seminar on Examinations organised by Ministry of Education, Govt of India at Bhopal* New Delhi Ministry of Education, Government of India, 1956

MINISTRY OF EDUCATION, GOVERNMENT OF INDIA (ED) *Education in Universities in India 1963-64* New Delhi Ministry of Education, Government of India, 1969

MISRA, G S *Prognostic Value of High School and Intermediate Marks* Unpublished M Ed Dissertation, Allahabad University, 1964

MISRA, V S Academic Achievement and Professional Success of Teachers Unpublished M Ed Dissertation, Allahabad University, 1964

MISRA, V S *An Investigation into Admission Criteria* Gauhati Examination Research Unit, Gauhati University, March 1968 (a)

MISRA, V S *A Study of Essay Type Test Reliability* Gauhati Examination Research Unit, Gauhati University, November 1968 (b)

MISRA, V S *A Study of the Effect of Randomisation and Scaling on the Errors in Examination Marks* Gauhati Examination Research Unit, Gauhati University, February 1969

MISRA, V S *A Critical Study of Essay Type Examinations* A Thesis presented for the Degree of Doctor of Philosophy, Gauhati University 1970 (Unpublished)

MITRA, S K A Factor Analysis of Examination Marks *Journal of Education and Psychology*, 1958, *16*, 382-402

MOORE, R A Methods of Examining Students in Medicine *Journal of Medical Education*, January 1954 Reprint cyclostyled by the Bureau of Educational Research, Allahabad

MUKERJEE, N *An Enquiry About Examination* Calcutta Das Gupta & Co, Ltd, 1954

MUKERJEE, N On Examinations *Journal of Gujarat Research Society*, 1961, *23*

MUKERJI, S N Improvement in Evaluation Procedures in Indian Universities Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon Madurai, 1970

MUKHERJEE, N M *External Examinations* Calcutta West Bengal Head Masters' Association, 1964, 338 343

M S UNIVERSITY, BARODA *Statistical Summary of Medical Examination* Report of the Students of the M S University of Baroda Baroda Statistical Service Unit, Faculty of Science, 1964

MYERS A E Simplex Structure in Grading of Essay Tests *Educational and Psychological Measurement*, 1966 *26*, 41 54

MYSORE UNIVERSITY *Report on University Examinations* Mysore Mysore University, 1965

NEDELESKY, L Exercise Writing in Natural Sciences Paper presented at the Invitational Conference on Testing Problems of Educational Testing Service, N J, 1957

NOLL, V H *Introduction to Educational Measurement* Boston Houghton Mifflin Company, 1957

ODELL, C W *How to Improve Classroom Testing* New York C Brown Company, 1958

PAGE, E B Grading Essays by Computers Progress Report *Psychological Abstracts*, 1967, *41*, 15820 (Abstracted from the Proceedings of the 1966 Invitational Conference on Testing Problems of Educational Testing Service, N J, 1967, 87-100)

PANDHARIPANDE, Place of Internal Assessment in Education *The Progress of Education*, 1962 *36*, 187-191

PATEL, I J Examination Reform Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon, Madurai 1970

PATEL, Y F Statistical Analysis of Marks Obtained at the B Ed Examination (1961) of S V Vidyapeeth *Studies in Education 1962 I*, 51-56

PENFORD, E. Symposium: The Use of Essays in Selection at 11+. Essay Marking Experiment: Shorter and Longer Essays. *British Journal of Educational Psychology*, 1956, *26*, 128-136.

PIDGEON, D. A. & YATES, A. Symposium: The Use of Essays in Selection at 11+: IV Experimental Enquiries into the Use of Essay-type English Papers. *British Journal of Educational Psychology*, 1957, *27*, 37-47.

PILLAI, N. P. *Report of An Investigation into the Organisational and Administrative Factors which Affect the Achievement of Pupils in Secondary Schools*. Trivandrum: The Department of Education, University of Kerala, 1965.

RAINA, T. N. Relationship between External Examination and Internal Assessment. *Teacher Education*, July 1963.

RAINA, T. N. Internal Assessment—How Much Dependable. *Education and Psychology Review*, July 1967.

RAM KRISHNA, V. A Scientific Basis of Examination Reform. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

RAO, T. S. M. Examination System in India: Need for Fresh Approach. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

RAY, ARUN. Reform of the Examination Technique. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

REMMERS, H. H. & Gage, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1955.

ROSS, C. C. *Measurement in Today's Schools*. Englewood Cliffs, N.J.: Prentice-Hall, 1947.

ROY, G. K. D. *Large Scale Failure in Mathematics—A Study*. Gauhati: Evaluation Unit, Board of Higher Secondary Education, Assam, 1969.

RYANS, G. DAVID. Measurement and Prediction of Teacher Effectiveness. Paper presented at the Invitational Conference on Testing Problems of Educational Testing Service, Princeton, N.J., 1958.

SARVATE, V. V. Pre-requisites for Examination Reform. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Madurai, 1970.

SECONDARY EDUCATION COMMISSION. *The Report of the Secondary Education Commission*. New Delhi: Govt. of India, 1952.

SECONDARY EDUCATION COMMITTEE. *Report of the Committee on Secondary Education*. Shillong: Assam Government Press, 1965

SHAH, R. M. Examination Reform. Paper presented at the tenth quinquennial conference of Inter-University Board of India and Ceylon, Mrai,adu 1970.

SHAPETI S B A Working Paper on Examination Reform Paper presented *at the tenth* quinquennial conference of Inter University Board of India and Ceylon Madurai 1970

SHARMA P D *Improving Examinations* New Delhi Directorate of Extension Programme for Secondary Education 1963

SIMS V M The Essay Examination Is a Projective Technique *Educational and Psychological Measurement* 1948 *8* 15 31

SINGH D P Reform in Examination System Paper presented at the tenth quinquennial conference of *Inter University Board of India and* Ceylon Madurai 1970

SIVAYYA K V Examination Reform Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon Madurai 1970

SNEDECOR G M *Statistical Methods* Ames Iowa Iowa State College Press 1940

SREENIVASAN M Examination Reform Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon Madurai 1970

S S C BOARD MAHARASHTRA *Statistical Investigations on Problems Pertaining to the Results of the S S C Examination* Poona S S C Board 1960

S S C BOARD MAHARASHTRA *Educational Evaluation and Testing Procedures* Poona S S C Board 1961

S S C BOARD MAHARASHTRA *A Statistical Analysis of Failures in English* Poona S S C Board 1963 (a)

S S C. BOARD MAHARASHTRA *Report of the Review Committee on S S C Examination Reform* Poona S S C Board 1963 (b)

STALNAKER J M The Essay Type Examination In E F Lindquist (Ed) *Educational Measurement* Washington D C American Council on Education 1951

STANDARDS COMMITTEE *Report on Standards of University Education* New Delhi University Grants Commission 1965

STARCH D & ELLIOT E C Reliability of Grading Work in Mathematics *School Review* 1913 254-259

STOREY ARTHUR G A Review of Evidence of the Case Against the True False Item *Journal of Educational Research* 59 6 February 1966 282 285

TAYLOR H J The Combination of Internal Assessment with Examination Results *Report on Examination Reform* U G C. 1962 Appendix X (a)

TAYLOR H J The Marks of Examiners *Report on Examination Reform* U G C 1962 Appendix XI (b)

TAYLOR, H. J. Classifying Students by Examinations. *Report on Examination Reform*. U.G.C., 1962, Appendix XII. (c).

TAYLOR, H. J. Choice and Chance in Conventional Examinations. *Report on Examination Reform*. U.G.C., 1962, Appendix XIII. (d).

TAYLOR, H. J. An Examination of Examiners: A Study of the Reliability of Examination Marking. *Journal of the University of Gauhati*. Science 1962, *13*. (e) also in *Journal of Educ. & Psychol.*, 1963, *21*, 15-25.

TAYLOR, H. J. *Supplementary Examinations*. Gauhati: Examination Research Unit, Gauhati University, 1962. (f)

TAYLOR, H. J., & TLUANGA, L. N. *Grace Marks: A Study of Criteria for the Adjustment of Examination Results*. Gauhati: Examination Research Unit, Gauhati University, March 1963. (a)

TAYLOR, H. J., & TLUANGA, L. N. *Scaling Tables*. Gauhati: Examination Research Unit, Gauhati University, 1963. (b)

TAYLOR, H. J. *Operation Passmark*. Gauhati: Examination Research Unit, Gauhati University, 1963. (a)

TAYLOR, H. J. The Question Paper. *General Education*, 1963, *4*, 1-9, (b)

TAYLOR, H. J. Mark Sheet Analysis. *The Journal of University Education*, 1963, II, 19-26. (c)

TAYLOR, H. J., & TLUANGA, L. N. *The Problem of the Third Examiner*. Gauhati: Examination Research Unit, Gauhati University, 1964.

TAYLOR, H. J. *Three Studies in Examination Technique*. New Delhi: University Grants Commission, 1964.

TAYLOR, H. J., & TLUANGA, L. N. *The Persistence Effect in Marking*. Gauhati: Examination Research Unit, Gauhati University, 1965.

TAYLOR, H. J., TLUANGA, L. N., & MISRA, V. S. *The Inconstant Examiner*. Gauhati: Examination Research Unit, Gauhati University, 1966. (a)

TAYLOR H. J., TLUANGA, L. N., & MISRA, V. S. *A Study in Multiple Marking*. Gauhati: Examination Research Unit, Gauhati University, 1966. (b)

TAYLOR H. J., TLUANGA, L. N., & MISRA, V. S. *Examination as Predictor*. Gauhati: Examination Research Unit, Gauhati University, 1966. (c)

TLUANGA, L. N., TAYLOR, H. J., TIEWSOH, F. N., ROY, & MEDHI, B. C. *The Influence of Scaling on Examination Results*. Gauhati: Examination Research Unit, Gauhati University, 1964.

THORNDIKE, R. L. *Personnel Selection. Test and Measurement Techniques*. New York: John Willey & Sons, 1949.

THORNDIKE, ROBERT L. Reliability. In E.F. Lindquist, (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.

THORNDIKE, R. L., & HAGEN, E. *Measurement and Evaluation in Psychology and Education*. New York: John Wiley & Sons, 1961.

THORNDIKE R L Reliability, 1963 In A Anastasi (Ed) *Testing Problems in Perspective* Washington D C American Council on Education 1968

TRAXLER A E Administering and Scoring the Objective Test In E F Lindquist (Ed) *Educational Measurement* Washington D C American Council on Education 1951

TRIVEDI H Examination Reform Paper presented at the tenth quinquennial conference of Inter University Board of India and Ceylon, Madurai 1970

UNIVERSITY EDUCATION COMMISSION *Report of the University Education Commission* New Delhi Government of India 1950

UNIVERSITY GRANTS COMMISSION (ED) *Evaluation in Higher Education* New Delhi University Grants Commission 1961

UNIVERSITY GRANTS COMMISSION (ED) *Examination Reform in Central Universities* New Delhi University Grants Commission 1969

VALIN E *The Value of Examinations* Paris UNESCO 1961

VERNON P E *The Measurement of Abilities* London University of London Press 1940

VICE CHANCELLORS CONFERENCE Comment on the First Draft Report of the University Grants Commission Committee on Examination Reform *Report on Examination Reform* New Delhi U G C. 1962 Appendix I

WISEMEN S The Use of Essays in Selection at 11+ III Reliability and Validity *British Journal of Educational Psychology* 1956 *26* 172 179

WOOD DOROTHY ADKINS *Test Construction The Development and Interpretation of Achievement Tests* Colombus Ohio Charles E Merril Books 1960

VARMA M The Invariance of the Standard Error of Measure *Journal of Education & Psychology* 1962, *20* 119 124

WOODS ROY C Can the Material Covered in Essay Tests Be Made Objective? *The Tenth Year Book of the National Council of Measurement Used in Education* Atlantic City N J 1953 80-90

WRIGHTSTONE J W JUSTMAN JOSEPH & ROBBINS IRVING *Evaluation in Modern Education* New York American Book Company 1956

Index

Tables and Figures are identified with numbers in italics, preceeded by *T* or *F*, thus *T172* means Table on or facing page 172, *F45* refers to figure on page 45. Proper names which appear only in the Review of the Literature (Chapter 5) have not been indexed

Abilities measured, essay vs objective examinations, 58-59, 62, 64, 80, 83

Abolition of J-effect, logic for, 40 194 n

Absolute differences of means, 158 159, *T170*, mean of, 162-163, *T170*

Academic aptitude test, used as link test, 249

Adding marks, illogicality of, 7, 35, 37, 160-161, 259

Agencies active in examination reform, 92

Aggregate marks reliability of, 7, 48, *T70*, 72, *T75*, 76, 78, 132n, weight of mathematics in, 162, 258

Agreement between examiners, examples of, 185, 201, 215, 233, on ranks vs on marks, 14-15, 255, self vs other (*see* Examiner agreement, self vs other) *See also* Examiner disagreement

Agreement per cent *See* Division, per cent agreement on

Agriculture, 78

Answer books preparation of, 100-102, 109-110, selection of, 99-100, 109

APA-AERA-NCME standards, 113

Aptitude tests, recommended, 91

Arabic examinations, 61

Arabic philosophy examinations, 61

Arts examinations, 69, 78, 158 *See also* particular subject

Assamese examinations, 67

"Average", inadequacy of, 20, 22, 124, 159, 163, 172, 185, 202, 216, 234

Average differences between examiners *See* Differences in marks between examiners, average

Average error, 5, 22 reduced by question-wise marking, 265 267 *See also* Errors in marking, Errors in marks

Average marks *See* Marks, mean

Average marks of disagreeing examiners more accurate than of agreeing 86-87, 260-264, *T263*

Averaging of marks unscientific, 49 *See also* Adding marks, illogicality of

Basumallik, T , 56-58, 114, 166

Bengali examinations, *T70*, 72, 76

Biology examinations, 6-7, 19 50, 67-71, 73 74, 76, 208 222, chance vs merit in, 213, choice of options in, determines marks, 220-221; compared with other examinations, 20 23, 35, 37, 41, *T70*, 157 177, 209, 216, 225 228, 229 n, 230, 234, different marks awarded equal ability in, 19 22, 32 35, 37, 220-221, difficulty of questions, variations in, 220-221, *T222*, distribution of marks, actual vs intended in, 161-162, Division in, changed on re-marking 24-32, *F30*, Divisions, distribution of, *T39*, 211-212 *T214*, Divisions, scatter charts of, 24-28, *T27*, 171, *T172*, 214, *T216*, 216-218, *T218* educational brinkmanship in, 209, Fail and Second to same students in, *T27*, item analysis of, 220-222, *T222*, largest differences between examiners in, *T21*, *T32*, 35, *T174*, 215, marks in, differences between, 19-20, *T21*, 158-160, *T214*, 214-216, marks in, distributions of, *F34*, 35-42 *F36*, *T38 39*, 208 209, *T210*, means of, *T39*, 158-160, *T170*, 210-211, 213, *T214*, 218-219, misclassification of students in, 20-22 71, 214; paper, description of, 208, 220, Pass Percentages in 6 22 24, *T23*, 212, *T214*, reliability, content, of, *T70*, 166-167, *T170*, 213 *T214*, *T288*, reliability, examiner, of, *T70* 164-165, *T170* 211-213, *T214*, *T216*, 216-218 *T218*, *T288*, reliability, marks, of, 42, *F44*, 46-48, *T70*, 168, *T170*, 213-214, *T214*, *T288*, representativeness of sample of, 100, scaling of, 209, 214, 219-220, "second" examiner in, 210,

self-consistency of examiners in, 218-220; self vs other disagreement of examiners in, 159-160, 163-164, 171, 217-218; six equal batches in, compared, *T32*, 32-35, *F33*, 171-175, *T174*, 209, *T210*, 217-219, *T218*, standard deviation of, *T39*, 41, 160-162, *T170*, 210-213, *T214*, 218-219; standard error of measurement of, 42-48, *F45*, *F46*, 170-171, *T170*, *T214*, 214, *T288*, standard of marking in, 41, 161, 211, 219, styles of marking in, 164-165, 213, 218; summary of statistics on, 4, 6-7, *T170*, 209-216, *T214*

Bose, P. K., 160

Brief summary: of Ninety Marking Ten, 4-6, 68, 74; of Four Thousand Re-examined, 4, 6-7, 68-69, 71, 73-74

Brinkmanship, educational, 40; in biology, 209; in Hindi, 194 n; logic for avoiding, 194 n; in mathematics, 224, naivite of, in marking, 194 n

Cambridge F.R.S and D.Sc., 61

Card punching, 104-105

Carelessness in original marking, 235, in second marking, 198

Central agency, need for, 94

Chance: cause of high failure rate, 13, 71; and pass-fail, 15-16, 70-71

Chance vs merit, 5, 7, 13, 15-16, 47; in biology, 213; controlled by scaling, 13, 16, 134, 243, 247, 258; in Hindi, 199; in history, 12, 130, 134, 139, 183, 186; in mathematics, 230, 241

Change, single easiest, 134, 242, 244, 259

Chemistry examinations, 67, *T70*, 71-72, 77

Chinese examinations, 54, 57

Choice of optional questions, 93; determines marks, 71, 84-85, 190-192, 220-221, 240-241; effect on reliability, 7, 8, 47-48, 83, 167, 182-183, 222, 230, effect on validity, 63-65, 85, 192, 240-241; guessing game, 192, 220, 241; reduction of effect of, 88; unjustified, 83 85, 241

Civics examinations, *T70*, 75

Class: in Tripos, 61 *See also* Division

Classes. *See* Divisions

Code used, 103-104

Coffman, W, 115

Commerce examinations, 66

Common batch: scaling method, 249; and standard deviation scaling, 251

Comparison of both history studies, 47-48, 167, 182

Computer, use of, 105-106

Consistency, measures of: absolute, 80-81, 116, *T126*, relative, 80, *T126*

Constant, adding or subtracting, as a scaling method, 252-253, 255

Content reliability *See* Reliability, content

Content, standard error of *See* Standard error of content defined

Contrast effect, 154

Conversion of marks, example of, 251-252

Conversion of rank orders to marks: example of, 256-257, methods for, 256

Conviction of infallibility in marking examinations, 2-3, 70

Correct mark. *See* Mark, correct

Correlation, between pairs of examiners (*see* Reliability, examiner); of examiners with criterion, 74, 150; of subjects at successive stages, 24, 78-79

Correlation coefficient, inadequacy of, 16-17, 80, 165, F169, 169-170, examples, 17, 148, 182, 198, 213, 229

Creativity, unreliably measured, 62, 64

Crosses and ticks, 67-68, 100-102, 109-110, effectiveness of crossing out, 102, 110, 285-287

Cumulative frequency curves, *F34*, 35-37

Cumulative grade point average, predicting with objective vs essay tests, 64

Cumulative records, recommended, 91

Cureton, E E., 150

Curriculum, improvement of, 81-82

Data: available for analysis, 103, 230, 289, reliability of, 99, 106-107, 111-112

Datta, S, 76, 132 n

D'Costa, A, 151

Definitions of reliability concepts, 113-127, *T126 See also* each concept

DEPSE, 65, 92, 123

Differences, absolute, 158-159, *T170*, mean of, *T170*

Differences in marks between examiners, 67-70, all four subjects compared, 19-20, *T21*, 158-161, *T170*, 172-175, *T288*, average, 6, 19 20, 22, in biology, *T214*, 214-216; examples of "agreement", 185, 201, 215, 233, in Hindi, *T200*, 200-202, in history, 8-10, *T9*, 14, 17, *T184*, 184-188, largest, 4-6, *T9*, *T21*, (*see also* largest differences), in mathematics, 229, *T232*, 232-234, measurement of, 248-249 *See also* Division; Divisions; Marks,

mean, Pass Percentages, Standard Deviation(s) (SD)

Difficulty index, 72

Difficulty of questions, paper setters cannot judge, 7, 82, 87, students cannot judge, 82, 192, 221, 240-241 *See also* Item analysis

Disagreement on best and poorest, 5, *T9*, 13-15, *T14*, 68

Disguising previous markings effectiveness of, 285-287, method, 100-102, 109 110

Distinction, compared with First, 225, and Fail to same candidate, 4-5, *T11*, 43, 129 130, 256, vs Third due to choice of question options, 240-241

Distributions of marks *See* Marks distribution of

Division accuracy of, 5-7, changes of, on re-marking, 6, 24-32, *T26 27*, *F29-30*, 67, dependence on chance vs merit, 4-5, 7, 32, 139, 243, disagreement on, 4-5, *T11*, 20, 24-28, 173, 243, meaning of, for history, 12, 132, minimum marks determined yearly, 88 untrustworthy, 11-12, 24 28, *F29 30* *See also* Divisions, scatter charts of

Division, per cent agreement on, 25, *T26 27*, 31, 67, 70, 185 186, 202 203, 216-218, 234-235, average, *T27*, compared with *r*, 187, 202-203, 217 218, 235 *For details see* Divisions, scatter charts of

Divisions awarded Rank 1 candidate, 139 140, and J-effect, 179, 194, 209, 224

Divisions, distributions of, 41-42, in all four subjects compared, *T39*, in biology, 211-212, *T214*, characteristic of examiner, 186, in Hindi, 197, *T200*, in history, *T11*, 132, *T133*, 180-182 *T184*, 186, in mathematics, 227-228, *T232*

Divisions, scatter charts of in all four subjects compared, 24-28 *T26 27*, 171, *T172* in biology, *T216*, 216-218, *T218*, in Hindi, *T202*, 202 203, *T204*, in history, *T186*, 186-189, *T188*, in mathematics, 234-237, *T236*, *T238*

D Sc , 61

Ebel, R L., 53-54, 59, 62, 72, 83, intraclass correlation, 122 123, 148, 166, 273

Economics examinations, 47 n, 69

Education examinations, 77, 79 80

Education Commission, 37, 49, 81, 85, 92, 271

Educational brinkmanship *See* Brinkmanship, educational

Einstein, A , 47 n

English examinations, 37, 60-65, 67-76, *T70*,79, 81, 243, more reliable than mathematics, 143, comparisons with history, 128, 136, 141-142, 150-151

Equal ability awarded different marks in all four subjects compared, 6, 19 22, 32-35, 37, 41, 158, in biology, 220-222, in history, 4-6, *T9*, 10, 130, 190-192, in mathematics, 240-241

Equating methods and rank order scaling, 255 *See also* Randomization and scaling

Equi percentile scaling, 250 *See also* Rank order scaling

Equivalent answer books, comparison of examiners on *See* Examiner X

Errors in marking, 239 n, average, 5, 22, not averaged out for failures, 48, not random, 10-11; in objective tests, 62, range of, 5, reduced in aggregate marks, 48, reduced by having different examiner for each question, 265 267, *T266*, reduced by multiple marking 5, 18, *T264* *See also* Standard error of marking

Errors in marks maximum possible, 10, 20-22, 130, mean, 119 120, minimum range of, 149, range of, 120-121, range of total, 126, theoretical distribution of, 282-283 *See also* Standard error of measurement (SEM)

Essay type examinations *See* Objective vs traditional examinations, Traditional-type examinations

Examination reform agencies active in, 92, publications for, 92-94

Examination research in India, handicaps to, 94 *See also* Research on examinations

Examinations (papers), selection of, 96-97, 108-109

Examiner agreement, self vs others, 6, 19, *T21*, *61*, in all four subjects, compared, 159-160, 163 164, 171; in biology, 217-218, in Hindi, 203, in history 187-188, in mathematics 19, 235-237

Examiner disagreement on "best" and "worst", 13 14, on Division, 11-12, examples of 'agreement', 185, 201, 215, 233, on marks, 8 10, on Pass Percentages, 12-13; on rank order, 13 14, on ranks and marks compared, 14 15, 255 *See also* under statistic sought

Examiner reliability *See* Reliability, examiner

Examiners carefulness of, 151; carelessness of, 198, 235; correlated pair with

different marks, example of, 17, differences between (*see* Differences in marks between examiners, *also* statistic sought), instructions to, 101, 104, 111; number needed, 262-264, *T263*, *T264*; original, in history, 149-151, *T151*, reliability of, when each marks a different question, 260, 264-267, *T266*, remuneration of, 104 111, sample studies of, 136 n, 146, *T147*, 150n, 151-154, selection of, 97, 111, training of, 61, 64, 123, 269, who agree likely to be wrong, 260-262, 264

Examiners, Indian as good as others, 7, 75, 151, 165, 246, reasons, 2

Examiners, self-consistency of in biology, 218-220, in Hindi, 204-205, in history, 188-189, in mathematics, 237-239

Examiner X (marking of six equal sets in each subject) all four subjects compared, *T32*, 32-35, *F33*, 172-175, *T174*; biology, 209, *T210*, *T214*, 217-219, *T218*, differences, *T32*, *F33*, *T174*, experimental design, 171-172, Hindi, 194, *T194*, *T200*, 203-205, *T204*, history, 178, *T178*, *T184*, 187-188, *T188*, 189, largest differences, *T32*, 172-173, *T174*; matching of sets, 103, mathematics, 224, *T224*, 230, *T232*, 235-236, 238, *T238*, and number of pages, 176-177; selection of, 98

Experimental design, 4, 8, 19, 96-112, answer books, preparation of, 100-102, 109-110, answer books, selection of, 8, 99-100, 109, card punching, 104-105, coding, 102-104, computation, 105-106; design, 98, 110-111, despatch and receipt, 104, 111, dividing, 102 104, examinations selected, 96-97, 108 109, examiners selected, 97, 111, examiner X, 98,171, formulas used, 106-107 (*see also* Formulas for), reliability of data, 99, 106-107, 111-112, reproduction, 108

Factor analysis of answer books, 151-152, of examiners, 152-154, of marks, 66, 73

Fail *See* Pass-fail line, Pass vs fail

Fail and First to same students in history, *T11*, physics, 67, mathematics, 67

Failure *See* Pass fail line, Pass vs fail

Failure rate, 66, caused by chance, 13, 15, 70, effect of randomization and scaling on, 13, 134 *See also* Pass vs fail, Pass Percentages

Findley, W, 113, 269

First Class, and F R S, 61

First Division: compared with Third, 5-6, *T11*, 20, 25, *T26-27*, 28-31, *F29-30*; in history failed, *T11*; in physics failed, 67; in mathematics failed, 67; and failure only reliable difference, 31. *See also* Division, changes of, on re-marking, Divisions, scatter charts of

Formulas for: average rank order correlation with criterion, 150, average rank order intercorrelation, 145; content reliability, 122 (*see also* Harper's content reliability formula), difference between correlated standard deviations, 106-107, marks reliability, 124; reliability concepts, *T126*, scales, relationship of, 114; standard deviation, Jenkins' estimate, 103n, standard deviation of differences, 148n, standard deviation of difference scores, 125; standard deviation scaling, 114, 251; standard deviation of a sum, 100, standard error of marking (unscaled), 125, *T126*; *t*-test, 106-107, 155-156

Four Thousand Re-examined; brief summary, 4, 6-7; detailed reports, 155-241; experimental design, 96-107, highlights, 19-50; summary tables, *T170*, *T172*, *T174*

Frequency distributions *See* Marks, distribution of

F R S, 61

Gauhati University *See* Taylor, *H.J*; Misra, V S

Gayen, A K, 2, *T70*, 79, 81-82, 85-90, 99, 167, 173 n, on choice of options, 85, 190, 220, 240; on mathematics, 230; on reliability, 71-74, 76

Geography examinations, 65, *T70*, 73, 75-76, 83, 93

Geometry examinations *See* Mathematics examinations

Grace marks, 7, 40, 49, 88, 93, 134; effect of Taylor's substitute, 279, percentage receiving, 99; Taylor's substitute recommended, 270, Taylor's substitute explained, 276-284

Grading categories for reliable marking, 7, on the curve, 160, 242, 244, recommended, 87

Guess, students forced to, 82

Guessing game, effect on marks, 192, 221, 240

Guilford, J P, 55, 76-77, 122, 136, 145, 148, 172 n

Gulliksen, H on reliability, 55, 57, 113, 115-116, 122, 124, 145 n, 165-166, 168; on scaling 46, 173-174 n, 246, 249

Half-marks, treatment of, in research design, *105*

"Halo" effect: and content reliability, 123, defined, 123, encouraged by Indian marking methods, 166

Harper, A E., *Jr*, *T70*, 83, *85-86*, 93, *115*, *on choice of optional questions*, *71*, 84, *109*, *on essay vs objective examinations*, *59*, 79, *on different marking systems*, *37*, *41*, *43*, *160-161*, on personal equation, calculation of, 239, 249, on scaling, *49-50*, *87-88*, *161*, 173-174 n, 242, 246, 249, *258*; on standard error of measurement, *80*, *170*

Harper's content reliability formula, 123, 166, 170, 273-275, effect of choice of options on, 199, 274-275, effect of differences in parts vs in questions on, 206, 241, effect of marking styles on, 183, 198-199, 213, 230, empirical check on, 124, example of absurd result from, 182

Head Examiner: better procedure, *86*, 247-249, *defines mean and standard deviation*, *252*, *role of*, *246-247*

Highest mark in sample underestimated, 173

Highlights. *See* Summary, highlights

Hill, W H., 59, 63, 81-85, 88, 90

Hindi examinations, 6-7, 19 50, 68-69, 71-74, 76, 160, 193-207, chance vs. merit in, 199; compared with other examinations, 23, 35, 41, *T70*, 157-177, 195, 203, 213-214, 226, 228, 229 n, 230 233-234, different marks awarded equal ability in, 19-22, 32-35; difficulty of questions, variation in, 206, *T206*; distribution, actual vs intended in, 161-162; Division in, changed on re-marking, 24-32, *F29*; Divisions, distributions of, *T39*, 197, *T200*, Divisions, *scatter charts of*, *24-28*, *T26*, 171, *T172*, *T202*, 202-203, *T204*, educational brinkmanship in, 194 n, Fail and Second to *same* students in, *T26*; item analysis of, 205-207, *T206*, largest differences between examiners in, *T21*, *T32*, 35, *T174*, 200-201; marks in, differences between, 19-20, *T21*, 158-160, *T200*, 200-202, marks in, distributions of, *F34*, *35-42*, *F36*, *T38-39*, 193-195, *T194*, means of, *T39*, 158-160, *T170*, 195-198, *T200*, 204-205; misclassification of students in, 20-22, 200; paper, description of, 193, 205, Pass marks in, 41, 194n, Pass Percentages in, 6, *T23*, 23-24, 197, *T200*; reliability, content, of, *T70*, 166-167, *T170*, 198-199, *T200*, *T288*, reliability, examiner of, *T70*, 164-165, *T170*, 196, 198, *T200*, *T202*, 202-203, *T204*, *T288*; reliability, marks of, 42, *F44*, 46-48, *T70*, *168*, *T170*, *199*, *T200*, *T288*, scaling of, 195, 200, 205; "second" examiner in, 196 *self-consistency of examiners in*, 204-205, *self vs. other disagreement of examiners in*, 159-160, 163-164, 171, 203; *six* equal batches in, compared, *T32*, 32-35, *F33*, 171-175, *T174*, 194, *T194*, *T200* 203-205, *T204*; *standard deviation of T39*, *41*, *160-162*, *T170*, 195-198, *T200*, 204-205, *standard error of measurement of*, 42-48, *F45*, *F46*, 170-171, *T170*, *T200*, 200, *T288*, *standard of marking in*, *41*, *161*, *197*, *205*, *styles of marking in*, *164-165*, 198-199, 203; *summary of statistics on*, *4*, *6-7*, *T170*, 195 202, *T200*

History examinations, 4, 8-50, pass in, 18, *37*, *65*, *68-69*, *71*, *73-76*, *128 154*, *168 169*, 178-192; both studies compared, 47-48, 167, 182, chance vs merit in, 12, 130, 134, 139, 183, 186, choice of options in, determines marks, 190-192, compared with English, 128, 135-136, 142, 150-151 (*see also* Taylor, H J, on multiple marking study compared with Ninety Marking Ten), compared with other examinations, 23, 35, 41, *T70*, 157-177, 213, 226, 228, 229 n, 230, 233-234, 236; compared with other studies, 68, *T70*, 108, 110, 113, 116, 128-129, 135-136, 140-142, 146n, 150-151, 154, 168; different marks awarded equal ability in, 4-6, *T9*, 10, 19 22, 32-35, 130, 190-192; difficulty of questions, variation in, 189-192, *T191*; distribution, actual vs. intended in, 161-162, Division in, changed on re-marking, 24-32, *F29*, Divisions, distributions of, *T11*, *T39*, 132, *T133*, 180-182, *T184*, 186, Divisions, scatter charts of, 24-28, *T26*, 171, *T172*, *T186*, 186-189, *T188*; Fail and First to same student in, *T11*; Fail and Second to same students in, *T26*, inter-examiner correlations in, *T147*; item analysis of, 189 192, *T191*, largest differences between examiners in, *4-5*, *T9*, *T21*, *T32*, *35*, *T174*, 184-185, marks in, differences between 8 10, *T9*, 17, 19-20, *T21*, 158-160, *T154*, 184-186, marks in distribution of, *9*, *F34*, 35-42, *F36*, *T38 39*, 129 131, *T130*, *T130*, *T178*, 178-179, *186*, *189*, *means of*, *T9*, *10*, *T39*, *69*, *T135*, 135-141, *T137*, *T142*, 143, 158-160, *T170*, 179-180, 182, *T184*, 188-189, misclassification of students in, 6, 10, 20-22, 130, 184; paper, *description of*, *178*, 190, Pass Percentages in, 6, *T12*, 12-13, 22 24, *T23*, 135-136, 141, 181-182, *T154*, reliability, content, of, *T70*, *T3*, 148-149, *166-167*, *T170*, 182-183, *T184*, *T288*, reliability, examiner, of, *16-17*, *26*, *T70*, 145-148, *T147*, 164-165, *T170*, 180-182, *T184*, *T186*, 186-188,

T188, 285, *T288*, reliability, marks, of, 42 *F44*, 47-48, *T70*, 149, 168, *T170*, 183, *T184*, *T288*, scaling of, 16-17, *T138*, 143, 179, 184, 189, 256-258, *T257*, "second" examiner in, 180, self-consistency of examiners in, 188 189, self vs other disagreement of examiners in, 159 160, 163-164, 171, 187 188 six equal batches in compared *T32*, 32 35, *F33*, 171-175, *T174*, 178, *T178*, 187-188, *T188*, 189, standard deviation of *T39*, 41, 135 141, *T138*, *T142*, 143, 160-162, *T170*, 179 182, *T184*, 188 189, standard error of measurement of, 16-17, 42-48, *F45*, *F46*, 141-144, *T142*, 170-171, *T170*, 183 184, *T184*, standard of marking in, 41, 101, 181, 189, styles of marking in, 164 165, 183, 186, 188 189 summary of statistics on, 4 7, *T170*, 179 186, *T184*

Hoyt and Stunkard, 122

Improvement of examinations *See* Traditional type examinations, improvement of

Indian Examiners, 2, 7, 75, 151, 165, 246

Indian research, 65 95, conclusions of, 90-91

Infallibility, conviction of, in marking examinations, 2-3, 70

Instructions to examiners 8, 19, 244-246, Head Examiner's, 247-248, improvement needed, 269, need for specific, 160, 165, 189, 205, 220, 236, recommended, 86

Intelligence tests, recommended, 91, 94

Inter-correlation of different academic levels, 78-79, of papers, 77

Internal assessment recommended, 89 91, scaling of, 89

Intra-class correlation, 122 123, 148, 166, 273

Item analysis in biology, 220-222, *T222*, and experimental design, 105, in Hindi, 205-207, *T206*, in history, 189 192, *T191*, in mathematics 239 241, *T240*, method used for, 190 205, 220 239, negative validity, 192, 206, 221, 240, should be done routinely for public examinations, 270

J-effect, 40-41, 86, 159 160 abolition of, logic for, 194 n, in biology, 209 210, and Division, 195, explanation of, 131, in Hindi, 194-196, in history, 41, 131, 179 180, in mathematics, 224-226, and need for scaling, 254 training examiners to eliminate, 40-41 *See also* Reverse-J

Jenkins' formula for estimating standard deviation, 103 n

Journalists, need for, 95, 270

Judgemental scaling, 238 239, 242, 244-246

Judgement in marking, 173, reduction of, 173

Justification for study, 1-2

Kelly, T L., on reliability standards, 47 n, 77

Keynes, John Maynard, 47 n

Kuder Richardson reliability, 122, 124

Languages, examinations in classical, *T70* *See also* under name of language

Languages, examinations in regional, 65, *T70*, 75 *See also* Assamese, Bengali, Hindi

Largest differences, 67 68 70, in all four subjects compared, *T21*, *T32*, 35, 173, *T174*, in biology, 215, in Hindi, 200-201, in history, 4-5, 79, 184-185, in mathematics, 6, 232-233, significance of, 172-173, in six equal sets, *T32*, underestimated 173 n

Law examinations, 64

Liddle, S K V , 158

Link test, 249, and standard deviation scaling, 251-252

Literature, review of, 53 95

Logic examinations, 69

Luck *See* Chance, Chance vs merit

Mahalanobis, P C , 37, 43, 46, 49-50, 87, 114, 160, 236, 242, 258

Maharashtra Board 65 66, *T70*, 75-76, 85 86, 89, 168

Margin of error *See* Standard error of measurement (SEM)

Mark, correct average of examiners, 10, 15, correlated with standard error of marking, 144, defined, 115 116, 120, distinguished from true mark, 115 116, 141, and grace marks 276-279, 282 283, in history, 140-141, *T142*, and original history examiners, 150, standard error of, 140

Mark, meaning of different for biology and mathematics, 41, 87, different for mathematics from English and history, 37, 243, different for mathematics from Hindi, 41, different for different candidates, 152 examiners 243 *See also* Scales used, Mathematics, scale different from other subjects

Mark, true defined, 42, 115 116, 170, distinguished from correct mark, 115-116, 141

Marking conviction of infallibility in, 2 3, on the curve', 160 242, 244, order of (*see* Order of marking first and second compared), standard of (*see* Standard of marking), standard error of, (*see* Standard error of marking), styles of (*see* Styles of marking), systems (*see* Scales used), variations in, 67-70, variations in, permissible limits of, 270-271

Marking methods average more accurate if examiners differ, 86-87, 260-264 *T263*, generally accepted criteria, use of, 62, impressionistic vs analytic, 60-61, 63, improvement of, 60-61 86-87 (*see also* Traditional type examinations improvement of), in India encourage "halo effect, 166, maximum marks determined by question difficulty 88, multiple marking 5, 18, 60-64, 86-87, 93, 260-267, question wise vs script wise marking 86 260-267, in science not objective, 69 227, three-stage, 248 249, two-stage, 245

Marks accuracy of 5 7, average, 5, 9, average related to range, 10, compared with ranks (*see* Marks and rank orders), determined by choice of questions, 190-192, 220-221, 240-241, differences between, (*see* Differences in marks between examiners) equal ability awarded different (*see* Equal ability awarded different marks), higher, to inferior candidate, 259, maximum for only part of syllabus, 85, median 140 *T142*, 144, of original examiners, 8, 149 151, *T151*, raw and scaled methods for converting 245 246, 249-252, reliability of, (*see* Reliability, marks), standard error of, (*see* Standard error of marks, Standard error of measurement (SEM)), standard error of combined (*see* Standard error of combined marks), validity of, 77

Marks distribution of actual vs intended 161-162, in all four subjects compared *F34* 35-42, *F36 T38 39*, in biology, 208-209 *T210*, characteristic of examiners 86, 136, 186, 189 219 effect of scaling on 257, in experimental design 99-100 102 103, 109 in Hindi 193-195, *T194*, in history, *T9* 129-131 *T130 F130*, *T178*, 178 179, 186 189 irregular so require normalized scaling 253-254, in mathematics 162 *T224* 224-225, 234, need to specify, 189, 205 219, 238 239, not normal 37, wide variety in, 37, 87 *See also* Divisions, distributions of, Specified distributions of marks

Marks, mean of all four subjects compared *T39*, 158 160 *T170* in biology, 210-211 213, *T214* 218-219, correlation of with standard deviations, 174, of disagreeing examiners more accurate than of agreeing 86-87, 260-264 *T263*, effect of differences in, on standard deviation of error, 171, of examiners compared with mean of correct marks, 141 150-151, *T151*, frequency distribution of, *T135*, 135 in Hindi 195 198, *T200*, 204 205 in history *T9* 10 69, *T135*, 135 *F137*, 141, *T142*, 143, 179-180, 182 *T184*, 188 189 in mathematics, 158 160 226-227 229 *T232* 234 236-238 range of, in other studies, 67-70 should be routinely calculated for public examinations 270 unrelated to range, 136-138 *T137*, used in scaling 251-253 255 variation in for six equal sets *F33*, 34-35, 171 175 *T174*

Marks range of awarded single answer book 5, *T9* 10, 67, 129, maximum, 129-130, minimum 129 related to ability 132, related to average, 10, related to ranks 140, scaled vs unscaled, 257 258 *See also* Range

Marks and rank orders differences between, 13 15 139-140 255 disagreement on, 14-15, 140, 255 and distribution of error, 144, table of original *T290*

Marks reliability *See* Reliability, marks

Matching of samples, 102 103

Mathematical methods of scaling, conditions for, 248

Mathematics examinations 6-7, 19-50, 47 n, 61, 66-76, 85, 223 241, chance vs merit in 230 234 choice of options determines marks in, 85 240-241, compared with other examinations 6 20 23-24 35, 19-50 passim *T70* 157-177 passim 163, 173 224-225, 227, 229 n, 233-234 236-237 different marks awarded equal ability in 6 19-22, 32-35, 37, 240-241, difficulty of questions variation in, 173, 240, *T240*, distribution actual vs intended in, 161 162, Division in, changed on remarking 24-32 *F30* Divisions distribution of, *T39*, 227 228 *T232* Divisions, scatter charts of 24 28, *T27*, 171, *T172*, *T236*, 234-237, *T235*, educational bunkmanship in 224 effect of optional questions in, 240-241, examination method defective in, 167 Fail and First to same student in 67, Fail and Second to same students in *T27*, improvement of, 167, item analysis of, 239-241, *T240*

largest differences between examiners in, 6, *T21*, *T32*, 35, 173, *T174*, 232-233, marks in, differences between, 6, 19-20, *T21*, 158-160, 229, *T232*, 232-234, marks in, distributions of, *T34*, 35-42, *F36*, *T38-39*, 162, *T224*, 224-225, 234, meaning of marks in, compared to other subjects, 37, 243, means of, *T39*, 158-160, *T170*, 226-227, 229, *T232*, 234, 236-238, misclassification of students in, 20-22, 71, 231, "objectivity" of, 69, 223, 227; paper, description of, 224, 239, Pass marks too low in, 27-28; Pass Percentages in, 6, 22-24, *T23*, 228 *T232*, reliability of, compared with history, 6, reliability, content, of, *T70*, 72-73, 166-167, *T170*, 229 230, *T232*, 234, *T288*, reliability, examiner, of, *T70*, 164-165, *T170*, 227, 229, *T232*, 234-237, *T236*, *T238*, 285 *T288*, reliability of, less than of English, 143, reliability, marks of, 6 42, *F44*, 46-48, *T70*, 168, *T170*, 230-231, *T232*, 234, *T288*, scale different from other subjects, 7, *F34*, *F36*, 37, 41, 43, 160 171, 173, 225, 237, 243 258, scale used in, 26-28, 31, 37 40, 225, 236-237, 258; scaling, need for, 231, 238-239, 'second' examiner in, 226; self-consistency of examiners in, 237-239; self vs other disagreement of examiners in, 19, 159-160, 163-164, 171, 235-237, six equal batches in, compared, *T32*, 32-35, *F33*, 171-175, *T174*, 224, *T224*, 230, 235-236, 238, *T238*, standard deviation of, *T39*, 41, 160-162, *T170* 226-229, *T232*, 234, 236-238, standard error of measurement of, 42-48, *F45*, *F46*, 170-171, *T170*, 231, *T232*, 234, *T288*, standard of marking in, 41, 161, 227-228, 236, 239, styles of marking in, 164-165, 230, 236, summary of statistics on, 4, 6-7, *T170*, 225-234, *T232*, weighted double/triple in aggregate, 162, 258

Maximum error *See* Errors in marks' maximum possible

McNemar, Q , 107

Mean errors, scaled and unscaled, 119-121; of marks (*see* Marks, mean), and standard deviation transformations, scaling by, 114, 249 254

Meaning of marks different in different subjects *See* Mark, meaning of

Measurement error: correcting grace marks for, 276-280, routine calculation recommended for public examinations, 270 *See also* Errors in marking, Standard error of measurement (SEM)

Medians' of correlation coefficients, 162, *T170*, of history marks, 140, *T142*, 144

Medicine, post-graduate examinations in, 59 n

Memory not a factor, 164

Merit, order of *See* Rank orders; Rank order scaling

Misclassification of students, *6*, *10*, 20-22, 70-71, 130, 184, 200, 214, 231

Misra, V S chapter by, 53-95; factor analysis by, 66, 73, on model answers, 86, on questions, 82-83, on reliability, *T70*, 73, 76, 115, 122, 124, 242, his reliabilities compared with ours, 285; research with Taylor and others, 68, 74, 78, 86-87, 128-129, on standard error of measurement, 80, on validity of Teacher Training courses, 79

Model answers' differences in standard not controlled by, 189, 219, recommended, 86, not supplied, 8, 19, 111

Multiple choice questions, recommended, 92

Multiple examinations, candidates faced with, 123, 138, 183, 199, 213, 230

Multiple marking, 74, 93, 260-267; effect of, 262 264, *T263*, *T264*, marking error reduced by, 5, 18, *T264*; number of examiners needed for, 262-264, *T263*, *T264*, question-wise marking compared with 265 267, reliability, effect on, 60-64, 86, 260, 266, by two vs three examiners, 86-87, 93 (*see also T263*, *T264*), validity, effect on, 63-64

NCERT, 85-86, 92 93

Ninety Marking Ten brief summary of, 4-6, 68, compared with Taylor et. al, 68, 108, 110, 113, 116, 128-129, 135-136; 140-142, 146n, 150-151, 154, detailed report of, 128-154; experimental design of, 107-112, highlights of, 8-18, selection of examiners for, 8, 108-109

Non-linearity *See* Item analysis, negative validity

Normalized scaling, 131, 250, 253-259

Objective examinations' main advantage of, 167, recommended, 85-86, 92, 268

Objective vs traditional examinations combined use recommended, 63; comparative reliabilities of, 6, 25, *56-58*, 166-168, 183, 199, 213-214, 230-231, comparative validities of, 61, 63-65, 79, conversion of traditional to objective, 58-59, 64, correlation with cumulative grade point average, 64, reliability concepts in, 47, 115-117, reliability of mark-

ing, 16, research on, 91, sampling, 167, standard errors of measurement of, 43, *F46*, validity of, in predicting essay marks, 63-65, 79, what is measured by? 58-59, 62, 64, 80, 83

Open-book examinations, recommended, 92

Optional questions *See* Choice of optional questions

Order of marking, first and second compared, 111, 129, *T130*, 132, *T133*, 139-140, *T140*, *T142*, 154

Order of merit, 2, 13-15, changed because marks unscaled, 161-162 *See also* Rank orders, Rank order scaling

Original examiners in history, 8, 149-151, *T151*

Other-agreement *See* Examiner agreement, self vs other

Oxford University, 61

Pages, number of, vs marks, 7, 49, 162, *T170*, 175-177, *T176*

Paper, description of · biology, 208, 220 ; Hindi, 193, 205 ; history, 178, 190 , mathematics, 224, 239

Papers . inter-correlation of, 77 , language of, 82 83 , selection of, 96-97, 108-109

Pass, compared with First Class, 5, *T11*, 25, *T26-27*, 28-31, *F29 30*, 67

Pass-fail line , in aggregate as unreliable as in individual subject, 48 , and chance, 15,70 ; effect of marking on 40 , unreliability of, 25, 68, 71, 134, 171

Pass vs fail , chance factors in, 15-16, 139-140 ; disagreement on, 5-6, 25, 32, 134 ; failures due to system, 134 , in matriculation vs P U , 79 *See also* Failure rate

Pass marks . in Hindi, 194 n , lowest in Asia, 194 n ; in mathematics, 27-28

Pass Percentages, 5-6, 160 ; in all four subjects compared, 22 24, *T23* , in biology, 212, *T214* ; compared with Taylor s study, 135 ; differences between examiners in, 22 24 *T23*, 68, 70, 90, 259 , in Hindi, 197, *T200* ; in history, *T12*, 12-13, 135-136, 141, 181-182, *T184* ; in mathematics, 228, *T232* ; as measures of teaching quality, 12, 22, 229 ; raised by randomization and scaling, 87, 134 ; representativeness of, in this study, 99 , of sum higher than of individual papers, 100 ; untrustworthy, 12 13, 22-24 ; variations in, 65 ; variations in, cause of, 12-13, 245

Passing probability, calculation of, 278-279, *T282*, 283

Percentage in each Division, routine, calculation recommended, 270

Per cent agreement *See* Division, per cent agreement on

Percentile ranks · changed because marks unscaled, 161-162 ; scale, 87, 254

Persistence effect, 83, 93, 110, 154

Personal equation, 239, 243 ; determining, 249

Physics examinations, 67, *T70*, 71-72, 77, 85, 93

Pooling of data distributions widened because of, 130 , effect of, on *r*, 26 n, *F169*, 169-170 , justification for, 24

"Popularity" of questions, 192, 221

Position effects, 154 *See also* Order of marking, first and second compared

Probabilities, multiplication of, 280, 283-284

Psychometrics/psychometry, nature of, 54

Publications for examination reform, 92 94

Public examinations · misclassification in, 6, 10, 20-22 70-71, 130, 184, 231 ; routine statistics to be calculated for, 270

Publicity, need for, 95, 270

Purpose of this research, 1-3, 8, 96

Questionnaire, administered to examiners, 286

Question papers *See* Paper, description of

Questions improvement of, 82 86 , increase in number of, recommended, 83 ; optional, (*see* Choice of optional questions), specificity needed in, 82-83 , types of, (*see* Paper, description of), vagueness of, lowers reliability, 82

Questions, variation in difficulty of, 93 , in biology, 220-221, *T222* ; effect of, on marks, 85, 190 220 240-241 , effect of, on reliability, 82 , in Hindi, 206, *T206* ; in history, 189-192, *T191*, in mathematics, 173, 240, *T240*

Question-wise marking, 264-267, *T266* ; cheaper and better, 260, 266-267 ; vs multiple marking 265-266 ; Radha Krishna's suggestion, 265 n

Qutub Minar effect, 40

Radha Krishna (V-C of Kanpur Univ) on question-wise marking 265 n

Radhakrishnan Commission. *See* University Commission Report (1950)

Rahman, H , 40

Ramanujan, 47 n

Randomization and scaling, 13, 93, 134, 238, 248 , alternatives to, 248-250 raises Pass Percentage, 134

Range defined, 129 n, 138 n ; estimation of SD from, 136, *T138* , for Examiner X, 173, *T174* , of marking error, defined, 120-121, *T126* , of marks awarded single answer book, 5, 9-10, 67, 129, *T257*, 257, of marks, scaled vs unscaled, 257-258 , of ranks awarded single answer book, *T14*, 14 , vs standard deviation, 172 n , of total error, 126 , unrelated to mean 136-138, *F137 See also* Marks, range of

Rank orders agreement on, would make scaling possible, 2 , average, for each answer book, *T14*, *T142* ; awarded each history answer book, 13-15, *T14*, 139-140, *T140* , compared with marks, 14-15, 140, *T142*, 255 , correlations of, *T147* , examiners disagree on, 13-14, 68 , mean of, 140, *T142*, 143-144 , range of, *T14*, 14 , standard deviation of, 140, *T142*, 143

Rank order scaling, 15, 87, 254-259 ; advantages of, 16, 144, 255-259 , effect on marks distribution of, *T257*

Rao, C R , 106

Raw marks, 245-246, 249-252

Reader reliability *See* Reliability, examiner

Recommendations, 81-95, 242-271 , determining marks reliability, 124-125 , elimination of optional questions, 83 85, 241, 269 , if examiners disagree, remark entire batch, 262 , grace marks, Taylor's substitute for, 134, 270, 276-284 , increase number of questions, 83, 269 ; instructions to examiners, 86, 269 ; miscellaneous, 268 , on multiple marking, 260-267 , multiple marking, when to appoint third examiner, 262 , outlaw J-effect, 40-41 ; question wise multiple marking, 260, 264-267 ; routine statistics to be calculated for every public examination, 270 , scaling, 69, 87 88, 134, 242-259 ; self-experimentation, 3, 269 , specifying marks distributions, 189, 205, 220, 238-239, 244-246, 250 ; substitute five or six categories for 100-marks scale, 7 ; training of examiners, 61, 64, 123, 269

Re-examination of answer books disadvantages of, 32 , failures would pass if, 71

Reform : consciousness for, 91-92 , single easiest, 242, 244, 259

Relative merit *See* Rank orders

Reliabilities, comparison of, using Spearman Brown formula, 57-58

Reliability (*major findings*), 2-3, 6-7 *F44*, *F45*, 47-48, 167-168, 174-175 , of biology examinations, 214 ; of Hindi examinations, 199 , of history examinations, 6, 8, 16-17, 134, 149, 151, 183 ; other Indian studies of, 77, 90 , of mathematics examinations, 6, 47, 231, 234 ; raised if each question marked by different examiner, 260, 266 *See also*, Choice of optional questions, effect on reliability , Reliability, content ; Reliability, examiner , Reliability, marks

Reliability (*see also* Reliability, major findings) · absolute measures of, 116, *T126* ; of aggregate marks, 7, 48, 72, *T75*, 76, 78-79, 132n , coefficient of, inadequacy of, 117, 169-170 ; concept applied to traditional examinations, 113-127 , defined, 55 , of examiners, factors involved in, 180, 196, 211, 227, function, of test length, 57 , index of, 150n, meaning of, in objective examinations 168 ; minimum required, 47, 76-77, 165, 270-271 ; practical utility of, 80-81 ; raised by scaling, 242 , reader, defined, 55 (*see also* Reliability, examiner) ; relative measures of, 118, *T126* , and scaling, 234 , total, defined, 55-56 (*see also* Reliability, marks)

Reliability, content, *T70*, 71-74, 117-118 ; for all four subjects compared, 165-167, *T170* ; for biology, *T70*, 213, *T214* ; comparable to reliability of objective tests, 57, 115, 166 , defined, 46-47, 55, 115-116 , effect of optional questions on, 7, 8, 47-48, 149, 167, 182, 222 , formula for, 122, *T126*, 273-275 , "halo" effect and, 123 , for Hindi, 198-199, *T200* , for history, 73, 148-149, 182-183, *T184* ; of link test, 249 , for mathematics, *T70*, 72-73, 229-230, *T232*, 234 , for mathematics compared with Gayen and Misra, 73-74, 230 , range of, 166 , routine calculation of, recommended for public examinations, 270 , table of, *T288* ; upper limits of, 166 ; variations in, explained, 73. 123, 166-167, 183, 198, 213, 230 *See also* Harper's reliability formula

Reliability, examiner, 60-62, 64, *T70*, 74-75, 117-118, for all four subjects compared, 163-165, *T170*, for biology, *T70*, 211-213, *T214*, *T216*, 216-218, *T218*, defined, 46, 114-115, differs from objective examination reliability, 114, 166, factors that lower, 86, highest found in Western studies, 62, for Hindi, 196, 198, *T200*, *T202*, 202-203, *T204*, for history, 16-17,

26, 145-148, *T147*, 180-182, *T184*, *T186*, 186-188, *T188*, for history compared with Misra's study, 285, increased by special training, 7, 166 (*see also* Traditional-type examinations), inter-examiner consistency (D) defined, 114-117, 122, *T126*, and marking criteria used, 62; for mathematics *T70*, 227, 229, *T232*, 234-237, *T236*, *T238* for mathematics compared with Misra's study, 285, methods for increasing, (*see Traditional type examinations, improvement of*), misleading if marks unscaled, 117-118, reduction in scripts won't improve, 69, routine calculation of, recommended for public examinations, 270, scaling more important than, 259, self-consistency (S) defined, 115 117, 122, *T126*, self vs other, 61; table of, *T288*

Reliability, marks, *T70*, 75-77, of all four subjects compared, 42, *F44*, 46-48, 167-168, *T170*, of biology, *T70*, 213-214, *T214*, definition and calculation of, 47, *115-116*, *124*, *T126*, *168*, *183*, *199*, *213*, 230, *of Hindi*, 199, *T200*, *of history*, 16-17, 144, 149, 183, *T184*; lowered for unscaled marks, 118, lowered by vague questions 82, of mathematics, *T70*, 230-231, *T232*, in objective examinations, 168, routine calculation of, recommended for public examinations, 270, table of, *T288*; of traditional examinations comparable to objective examination reliability, 47, 57, *115*, 166, *168*

Reliability of marking. *See* Reliability, examiner

Reliability, total *See* Reliability, marks

Reproduction, methods of, 108, 110

Research design *See Experimental design*

Research on examinations earliest, 54, efficiency aspects 66-81, foreign, *58-65*, functional aspects, 65-66, handicaps to, 94, improvement, *81-90*, Indian, 65-91 and this book, interpretation of, 95, needed, 94-95; summary, 90-91

Reverse-J, *in biology*, 209, *in Hindi*, 195, *in history*, 131, 179, in mathematics, 225 *See also J-effect*

Review of the literature, 53 95

Rotaprint, 108

"Same" vs. "Different" re-examiner *See* Examiner agreement, self vs. other

Sample adequately representative, 99-100, 109, 136, of five hundred adequate for routine examination statistics, 270, study of stratified random, 136 n, 146, 146 n, *T147*, 150n, 151-154 *See also Experimental design*

Samples, matching of, *102 103*

Sampling in objective vs. traditional examinations, *167*

Sanskrit examinations, *T70*, 72, 75

Scaled marks. *See* Scaling methods raw vs scaled marks

Scales, converting *non linear to linear* 255 *See also* Rank order scaling

Scales used, 7, 41; adding different, 7, *35*, *259*, for biology, 209, effect on Pass Percentage, *13*, 245, effect on standard error of measurement, *171*; examiners differ in, 136-139, for Hindi, 195, for history, 16-17, *T138*, 143, 179, for mathematics, 26-28, 31, 37, 40, 225, 236-237, 258, for mathematics and for biology compared, 41, 237, for mathematics different from other subjects 7, *F34*, *F36*, 37, 41, 43, 160, 171, 173, 225, 237, 243, 258 for mathematics and for English compared, 37, 243, reason for wide variety in, 37, 244, for science marks higher than for arts, *158*

Scaling 2, 242 259, defined, 114, 243; *of different subjects to common standard*, 258 259, effect on Pass Percentages, 13, *87*, 134, effect of, example, *251-252*, *T257*, *257-258*, effect of not, 17, 148, 161, 258-259, examiners, 242, 244, example of, 251-252, *T257*, formula (*see* Scaling methods) importance of, more than of examiner reliability, 259, of individual questions, 87, 221, 241; of mathematics, 27 28, 37-40 randomization and, *13*, *93*, 134, *255*, *248-250*, recommended, 69, *87-88*, 134, 242 259, reduces standard error of measurement, *46*, *88*, and reliability, *81*, *117*, *146-147*, 234, 242, simplicity of 242, single raw-set change, 242, 244, 259 and standard errors, 118 121, subjects 242, 244, 258-259

Scaling need for, 49 50 69 87 88, 251, 258, in biology, 214, 219-220 in Hindi, 200 205, in history, 184 189, in mathematics, 231, 238-239 *See also* Scaling

Scaling methods, 93, 244-259, by adding or subtracting a constant, 252 253, common batch, 249, 251; compared with temperature scale conversion 7, 27, 49-50, 114, 173-174 243, described, 250, equi-percentile, 250, formula *71*, effect of, 139, "grading on the curve", *160*, 242, 244, judgmental 238 253 242, 244 246, *kick test*, 247 251, mathematical methods of, cited uses for, 248, by mean of marks, 251 253 253, mean and standard deviation transformation, 114, 131, 242-254, normalized, 131, 249, 131, 242-254, normalized, 131, 249,

253 259, randomization and scaling, 13, 93, 134 238, 248 250, random sample adequate for, 249 rank order scaling, 15 16 87, 144, 254-259, *T257*, raw vs scaled marks 245 246, 249 252, re-examining small samples 247-248 251, role of Head Examiner, 246-248, simplicity of, 242, specifying distributions of marks 189 205, 219, 238 239 244-246, 250, standard deviation scaling, 114, 131, 249 254, standard scores, 250, use of selected samples in, 247, 249, without randomizing scripts, 248 249

Scatter charts of Divisions *See* Divisions, scatter charts of

Scholastic aptitude tests, recommended, 91

Science examinations, 65 69 70, *T70*, 75, 78 85, 93 158 *See also* Biology, Chemistry, Mathematics, Medicine, post graduate, Physics

Score reliability *See* Reliability, marks

Score, true, 170 *See also* Mark, true

SD *See* Standard deviation(s) (SD)

Second Divisioners awarded Distinction and First to Fail *T9*, *T26 27*, *F29 30*, 43

'Second' examiner, consistent features of in biology, 210, in Hindi 196, in history, 180, in mathematics, 226

Self agreement *See* Examiner agreement, self vs other, Examiners, self-consistency of

Self-consistency *See* Examiners, self consistency of

Self-correlation *See* Reliability, Examiner

Self-experimentation recommended, 3, 269

SEM *See* Standard error of measurement (SEM)

Semester system recommended, 90

Short answer examinations recommended, 92 reliability and validity of, 269

Significance, statistical vs educational, 159, 164 n 177

Six equal batches marking of *See* Examiner X

Skewness, 144

Snedecor, G M , 136, 138, 146

Social studies examinations, *T70*, 75

Spearman Brown Prophecy formula 57 58

Specified distributions of marks, 244-246, 250, method, 244-246, need to specify, 189, 205, 219, 238-239

Srivastava, A B L., 107

Standard deviation(s) (SD) of all four subjects compared, *T39*, 41, 160-162, *T170*, of biology, 210-213, *T214*, 218-219, correlation with means, 174, determines weight of paper, 160-162, 253 258, estimated from range, 136, *T138*, effect of, on standard error of measurement, 171, of Examiner X, *F33*, 34-35 172 174, *T174*, of Hindi, 195, 198, *T200* 204-205, of history, 135-141, *T138*, *T142*, 143, 179 182, *T184*, 188 189, of history correct marks, *T142*, of mathematics, 226-229, *T232*, 236-238, vs. range, 172 n, routine calculation of, recommended for public examinations 270, used in scaling (*see* Standard deviation scaling) variations in, in six equal sets, *F33*, 34-35

Standard deviation scaling, 114, 249 254, advantages of, 253, disadvantages of, *114*, *131*, *254 examples of*, *252*, *formulas for*, 114, 251

Standard error of combined marks, *T126*

Standard error of content defined, 116, *T126*

Standard error of estimate, more stable than *r*, 170

Standard error of estimation, 116 n *See also* Standard error of marking

Standard error of marking defined, 116, *example of*, *148*, *formulas for*, *125*, *T126*, in history, 141 142, *T142*, 144, 148, scaled (S) and scaled (D) defined, 118-121, 125, *T126*, and standard deviation, 141, unscaled (D) defined, 119 121, 125, *T126*

Standard error of marks, 116-118, 121, 125 126, *T126*

Standard error of measurement (SEM), 116-118, 121, for all four subjects compared, 42-48, *F45*, *F46*, 170-171, *T170*, absolute consistency, measure of, 80-81, *116*, *actual errors larger than*, *48*, *171*, best indicator of reliability, 80-81, 116, 170, for biology, 214 *T214*, correcting grace marks for, 277 280, defined, 42, for Hindi, 200, *T200*, for history, 16-17, 149 183 184, *T184*, interpretation of, 42 170 largest for mathematics, 43-46 *F45*, *F46*, 171, for mathematics, 231, *T232*, 234, minimum estimates of, 171, for objective examinations, *F46*, reduced by scaling 46, table of, *T288*, theoretical distribution of errors, 282-283 *See also* Standard error of content defined, Standard error of marking, Standard error of marks

Standard of marking 90, 173 174, in all four subjects compared, 35, 161, average 159 160, in biology, 211, 219, constant characteristic of examiner, 86,

189, 219, and Head Examiner, 246-247, in Hindi, 197, 205, in history, 181, 189, in mathematics, 227-228, 236, 239, in this study, 111, variations in 86, 90, 205, 219

Standards, APA-AERA-NCME, 113

Standards for reliability, 47, 76-77, 165

Standard scores, 250

Stanines, 250, 254

Statistics, routine, calculation for every examination recommended, 270, sample of 500 adequate for, 270

Studies compared, both history, 47-48, 167, 182

Styles of marking, 164-165 in biology, 213, 218, effect on reliability, 73 115, 123, 166, 183, 198 199, 213, 230, in Hindi, 198-199, 203, in history, 183, 186, 188 189, in mathematics, 230, 236 *See also* Standard of marking

Subjects studied, 4, selection of, 96-97, 108 109

Summary brief, 4-7, 68 69, 71, 73 74, highlights, 8 18, 19 50, of reliability concepts, *T126*, table explained, 157, tabular, *T70*, 157-171, *T170*

Summary of other Indian research, *T70*, 90-91

Supplementary examinations abolition of, recommended 88 89

Taylor, H.J , 2, *T70*, 85, 88 89, on adding marks, 37, on cause of high failure rate, 13, on failure and chance in public examinations, 13 15 on grace marks, 40, 49, 88 134, 270, 276-284, on J-effect, 40, 86, 131, on matriculation vs P U , 78-79, on multiple marking (reliability), 74, on multiple marking study compared with Ninety Marking Ten 68, 108, 110 113, 116 128 129, 135 136 140-142, 146 n 150-151, 154, on optional questions, 83 84, on Pass Percentage, 13 on randomization and scaling, effect of, 13, 134, on re-marking equivalent scripts 69-70, on re marking same scripts 69, on scaling, 49 69, 87 88, 134 236 248, 258 259, on third examiner, 86-87, 93, 261, 264, on third examiner, his argument simplified 261 262, on validity, 77

Teachers, training of, 95, 269, validity of, 79 80

Temperature scales as examples of scaling, 7, 27 49 50, 114 173 174, 243, conversion formula 243

Terms which are used interchangeably, 56

Third and First Divisions to same students, 6, *T11*, 20, 24-25, *T26 27*, *F29 30*, in mathematics, *T27*, *67*

Third examiner question-wise marking better than 266-267, recommendation on when to appoint, 262, Taylor s argument on, simplified, 261 262, unjustified when two examiners differ, 262 264

Third examiner, average of two more accurate if differ, 86-87, 93, 260-264, illustrated, 261 264, *T263*

Thorndike, R.L , 58 80, 83, 116, 118, on reliability standards, 47 n

Tick marks, 67 68, disguising of, 100-102, 109-110, effect of, 102, 110, 285 287

Total reliability *See* Reliability, marks

Traditional type examinations combined with objective 63, compared with objective, 6, 25, 43 *F46*, 56-58 79,91, 166-168, conversion of, to objective items, 58-59, 64, defined 113 114, improvement of, 2 3, 58, 60-61, 81 95, 194 n, 236 268 270, 285 (*see also* Recommendations, Scaling methods), instructions to examiners need improvement, 269, poor reliability and validity of, 2 60-64, 77 80, 90-91, 94 268, reliability, calculation of, 273 275, training of examiners needed, 269, weakness of, 61, what is measured by ?, 58 59, 62, 64, 80, 83, 269 *See also*, Objective vs traditional examinations

Tripos, Class in, 61

True mark, 42, 115 116, 170

True score *See* True mark

T-scale, 250, 254

t test inapplicability of, 106, note on use of, 155 156

University Commission Report (1950), 81, 85, 91, 268

Unscaled marks effect of adding 161, and inferior candidate, 259

Validity of B Ed examinations, 77, can be higher than reliability, 168-169, defined, 56, effect of optional questions on 63 65 192, 240-241, of objective vs traditional examinations, 61, 63 65, 79;

practical utility of, 80; of science marks, 77; of Teacher Training courses, 79-80; of traditional examinations, 60-64, 77-80, 90-91, 94; of class in Tripos, 61

Variations in marking, other Indian research on: conclusions about, 90; in equivalent samples, 69-71; in same scripts, 67-69

Varma, M., 170

Verbal intelligence test, used as link test, 249

Vernon, P., 58, 63, 76, 262

Viva voce: essay test recommended with, 9; recommended, 91

Wastage and stagnation, 66, 90

Writing ability, validity of essay vs. objective tests in predicting, 63-64, 79

X. *See* Examiner X

Xerox, 108

Zero not a measurement, 225